The symmetric eigenvalue problem: stochastic perturbation theory and some network applications.

Stoyanov, Zhivko

Award date:
2008

Awarding institution:
University of Bath

The symmetric eigenvalue problem: stochastic perturbation theory and some network applications.

submitted by

Zhivko Stoyanov

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

October 2008

Summary

This thesis is concerned with stochastic perturbation theory of the symmetric eigenvalue problem. In particular, we provide results about the probability of interchanges in the ordering of the eigenvalues and changes in the eigenvectors of symmetric matrices subject to stochastic perturbations. In this analysis we use a novel combination of traditional Numerical Linear Algebra, Perturbation Theory and Probability Theory. The motivation for this study arises from reliability of spectral clustering of networks, when network data is subject to noise. As far as we are aware, there is nothing comparable in the literature.

Further, we make conjectures from which we derive an asymptotic relation between the distributions of the largest eigenvalue and the 2-norm of random symmetric matrices, whose entries above the main diagonal are independent, identically distributed random variables with probability density functions being symmetric with respect to zero, including matrices from the Gaussian Orthogonal Ensemble (GOE). As far as we know, some of these conjectures are not new (possibly only as conjectures) but we are not aware of any proofs.

Also, we consider networks of coupled oscillators. In their analysis we use both, knowledge of dynamical systems and spectral properties of non-negative matrices. As a result, we present an algorithm, which uncovers the "master-slave" structure of the network. With its help, the analysis of the dynamics and the entrainment of the entire network can be reduced to considering only few of the oscillators, those whose dynamics determine the behaviour of the rest. This can be helpful in large networks exhibiting the "master-slave" structure.

Finally, we consider similarities of spectral clustering with respect to different matrices which can be associated with a given network. In particular, we compare clustering of products of Path graphs with respect to two different matrices: the Laplacian and the Normalised Laplacian matrices of the graph. We make the comparison by constructing a Homotopy between two eigenvalue problems and, using some Linear Algebra techniques, we show that the two matrices give similar spectral clusterings when applied to products of Path graphs.

Acknowledgements

I was fortunate to have so many good people as friends, who contributed in one way or the other to the writing of this thesis.

Alastair Spence, my supervisor, is certainly the one who has suffered the most throughout my academic, but also emotional, development as a Ph.D. student. However, he never gave up on me, even when things didn't seem as shiny as they do today! Thank you Alastair, not only for showing me how to do Maths, but also for being a true friend!

Many thanks to Des Higham, Gabriela Kalna, Keith Vass and Peter Grindrod for all the enjoyable meetings in Bath and in Glasgow! Des could always explain pages of subtle Maths with only a couple of words and Pete would make you laugh with every story he tells, and they were many!

Geoff Smith, who will probably never read this thesis, has shown me what beautiful Maths really is, by inviting me to all the IMO camps. His joy, when solving Maths problems, and his humour have always been contagious.

Alan Zinober, my former supervisor from Sheffield, was the one who gave me the hint that Alastair and Bath would probably be the right choice for me and I am glad I followed his advise. Thank you Alan, for being a wonderful friend and for making me laugh to the point of tears!

Many thanks to my friends and colleagues from the University of Bath for all the good moments spent together! In particular, thanks to my officemates throughout the years, Ant, Phil, CF, Ray, Bruce, Melina, Simone, Dave and Laura, for making the time at work a pleasure, and to all those who work hard behind the scene to keep the Department ticking, Ann Linfield, Carole Negre, Jill Parker, Mary Baines, Sue Paddock, Eric Wing, Steph Skaife and Sarah Hardy! Also, special thanks to Mark, Darrel, Adam and Jess from computer support, who still manage to help everyone despite all odds!

Many thanks to Yana, Sophie and Ivan for being so nice and helpful to us all the time!

My heartfelt thanks to Mum and Dad for giving me all their love, since I was far too little to know what Maths is! Also, a big hug to my brother, Svilen, who was strong enough to be the younger of us both!

Last, but far from least, I thank my family, Ivelina, Viktor and Boris, from the bottom of my heart! They have always been there for me, supporting me with their quiet strength!

This work was supported financially by grant number GR/S62390/01, provided by the EPSRC.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Motivation.

Networks are becoming an increasingly important area in mathematics, computer science and physics because of their many applications in, for example, internet, search engines, social networks, biological networks, vaccination strategies, airline hubs/nodes, etc. We briefly describe only few of them. For an informal treatment of the recent applications of networks see (Barabasi, 2003).

Network theory nowadays is used in finding effective, yet cost-efficient ways of vaccinating people, by examining the network of social interactions between them and vaccinating only the most significant *hubs*, that is, the people with the largest number of acquaintances. In order to do that, network scientists are searching for models, which are susceptible to scientific analysis and, at the same time, represent good resemblance with the real network of interactions between the people. For example, it has been discovered that, despite the large number of people living on the planet, approximately 6 billion, the average distance between any two people is six other people. This discovery, and many others, have confirmed the assumption that social networks can be modelled by the so-called *scale-free* networks (see (Barabasi, 2003)).

In search engines (e.g. Google) the web pages and the links between them are considered as nodes and edges in a network. Based on this, the task of a search engine is to rank the pages according to some criteria, for example, the number of pages linking to a given page, relevance of its contents, reliability, etc.

The world wide web, by its structure, is similar to a social network. With approximately 800 million web pages, the average distance between any two web pages is only 19 clicks. Recent advances in computer technology and the increasing volume of the world wide web, together with its good features, has caused many concerns, as we now describe. For example, the resilience of a given network, or a sub network, to an attack and the possible global and local impacts in case of a failure of that network. As an interesting fact we note that, by considering network models, it has been discovered that the world wide web is in fact a very "robust" network. In particular, if 5% of the nodes

on the web are removed at random, the communication between the remaining nodes will remain unaffected. However, if 5% of the most connected nodes on the web are removed, that is, if the web is subject to an attack, then the average distance between any two web pages doubles, becoming twice 38 clicks. In this respect, the analysis in this thesis is believed to be a first step at providing analytical tools for understanding the sensitivity of networks to disturbances.

In analysing micro-array data, networks are used to uncover relations between different genes, which helps scientists to understand certain diseases better and improve, or possibly invent new treatments. For example in cancer research, by considering samples of healthy and cancerous tissues and by clustering the genes of these tissues into groups, according to their activity, one is able to distinguish new types of cancer or differentiate between existing types. Because of the large number of genes present in each tissue, conventional discrete methods for clustering become impractical. Instead, clustering with respect to an eigenvector (or singular vector) of a matrix associated with the micro-array data is used. This is called *spectral clustering*, due to the fact that clustering is in fact done with respect to a part of the spectra of the matrix. In particular, micro-array data can be represented as a network by thinking of the tissues and the genes as nodes, where the tissues form a separate group from the genes. The links between the nodes in that network are only between genes and tissues, with no links connecting genes with genes, or tissues with tissues. The strength (or *weight*) of a link between a tissue and a gene is measured by the activity of the gene in that particular tissue. This type of network, consisting of two groups of nodes with no links between the nodes in the same group, is called a *bipartite graph*.

While efficient in terms of computational cost and time, spectral clustering is a heuristic technique and only approximates the solution to the discrete problem of clustering. Ideally, the clustering obtained by spectral methods is very close to the real (discrete) solution, but in some cases it could be very poor (c.f. (Guattery and Miller, 1998)). This is why many authors have designed techniques of spectral clustering aimed at specific applications, for which the technique is "known" to work. This is particularly true for the spectral methods of clustering micro-array data. We shall not aim to list these methods here, because the spectral clustering of micro-array data, as such, is not the main subject of this thesis. However, the following paper (c.f. Higham et al. (2005)) contains the point of departure for our work. Specifically, the goal of our research is to provide means for measuring the sensitivity of spectral clustering to perturbations in the network. In the context of micro-array data, which is typically very noisy, a potential application of our research would be to predict possible misclassification of certain genes, or genes, in general which are due to the noise in the data.

Since spectral clustering is done by considering parts of the spectra of a matrix associated with the network, the analysis of sensitivity of spectral clustering to noise in the data can naturally be stated as the problem of perturbation of the eigenvalues and the eigenvectors of matrices. In particular, a major aspect of this thesis is concerned with the sensitivity to perturbation of the spectra of symmetric matrices, which usually correspond to networks in which the links between the nodes have no directions, that is, if node a is linked to node b, then b is also linked to a. While, for example, the world wide web is not an example of such a network, the micro-array data and the social contacts between people can be considered as undirected networks. In this respect, the feature of our work, which we consider new, is that we analyse random perturbations to matrices, which correspond to random perturbations in networks. This, we believe, is a better representation of the noise, occurring in networks in reality. Mathematically, this is represented as a perturbation of a deterministic symmetric matrix by a random symmetric matrix multiplied by a scalar parameter which represents the magnitude of the perturbation (noise). The random matrices, by which we perturb, are mostly scaled matrices from the Gaussian Orthogonal Ensemble, but the settings for most of these results are in fact quite general. This allows for possible extensions to the theory here, for better models of particular types of random or deterministic noise.

There are methods available in the literature, e.g. (Newman et al., 2008), which address the problem of random perturbations to networks and the corresponding sensitivity of the spectral analysis (clustering or modularity) of these networks. Some of these methods are indeed very sophisticated in finding adequate measures of the similarities (or distances) between different clusterings, but their implementation is based on simulating the random perturbation. This can be very costly computationally, when the network is very large. This is why our efforts in this thesis have been directed towards developing flexible analytical tools which are computationally cheaper than performing simulations.

1.2. Background Graph Theory.

1.2.1. General definitions and notation in Graph Theory.

It is convenient to revise some standard notation in Graph Theory, which we use mainly in §6.

Definition 1.2.1 (**Graph**). A graph, G, is a collection of nodes, V_G, together with a set of links, E_G, between those nodes. Formally, *graph* is an ordered pair $G = (V_G, E_G)$, where V_G is a set of *vertices* and $E_G \subset V_G \times V_G$ is a set of relations between those

vertices. Usually, the elements of the set E_G are called *edges*.

Definition 1.2.2 (**Order of a graph**)**.** Let $G = (V_G, E_G)$ be a graph. We say that G is a *finite graph* if the set V_G is finite. Also, we say that G is a graph of *order* n, $n \in \mathbb{N}$, if V_G contains exactly n elements.

Definition 1.2.3 (**Undirected graph**)**.** We say that the graph $G = (V_G, E_G)$ is *undirected*, if the set of relations (*edges*), E_G, is symmetric, that is, if $(u, v) \in E_G$ implies $(v, u) \in E_G$ for all $u, v \in V_G$.

Definition 1.2.4. We say that the graph $G = (V_G, E_G)$ contains *a loop*, if $\{v, v\} \in E_G$ for some $v \in V_G$.

Remark 1.2.1. *In this thesis, unless otherwise stated, the graphs that we consider are* undirected *and without* loops*. Therefore, following some accepted notation in Set Theory, we shall denote an edge between vertices u and v by $\{u, v\}$, instead of (u, v), since the latter is mostly used to denote ordered pairs. We shall reserve the notation (u, v) for later, to denote vertices in cartesian products of graphs, where the order of appearing of u and v does matter.*

Definition 1.2.5. Let $G = (V_G, E_G)$ be a graph. We say that the vertices $u, v \in V_G$ are *adjacent*, if the pair $\{u, v\} \in E_G$. Sometimes we shall denote that u is *adjacent* to v by $u \sim v$.

Definition 1.2.6 (**Weight function**)**.** Let $G = (V_G, E_G)$ be a graph and $W = V_G \cup (V_G \times V_G)$. *Weight function* on G is any function $w : W \to \mathbb{R}_{\geq 0}$ such that $w(v) \geq 0$ for all $v \in V_G$, $w(u, v) > 0$ for all $\{u, v\} \in E_G$ and $w(u, v) = 0$ otherwise.

Remark 1.2.2. *In Definition 1.2.6 and in §6 we use the notation $w(u, v)$, instead of the correct but clumsy $w\big(\{u, v\}\big)$, to denote the* weight *of the edge $\{u, v\}$.*

The "usual" weight function, associated with a given graph $G = (V_G, E_G)$, is

$$w(u, v) = \begin{cases} 1 & \text{if } u \sim v; \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

The weight function w, defined by (1.1), treats all edges in a similar way, by giving them equal weights.

Definition 1.2.7. Let $G = (V_G, E_G)$ be a graph and w be some weight function on G. We say that w is *consistent*, if

$$w(v) = \sum_{u \in V_G} w(u, v).$$

Remark 1.2.3. *When w is a consistent weight function, the quantity $w(v)$, $v \in V_G$, is usually called the* degree *of vertex v and is denoted by d_v. So in this thesis, unless otherwise stated, we shall consider graphs with* consistent *weight functions on them and thus, we shall refer to the weights of the vertices of the graph as their* degrees*. This is helpful in destinguishing between weight function on edges and weight function on vertices, since the former requires two vertices as arguments and the latter only one - sometimes this could be confusing.*

Definition 1.2.8. A weight function which is not consistent is called a *non-consistent* weight function.

Definition 1.2.9 (**Unweighted graphs**)**.** Graphs $G = (V_G, E_G)$ with weight functions w defined by (1.1) are called *unweighted graphs*.

Definition 1.2.10 (**Weighted graphs**)**.** Graphs with a general weight function are called *weighted graphs*.

Definition 1.2.11 (**Connected graph**)**.** Let $G = (V_G, E_G)$ be a graph. We say that G is *connected* if every pair of vertices can be connected by a path along non-zero weighted edges.

1.2.2. Laplacian matrices of graphs

Definition 1.2.12 (**Laplacian matrix of a graph**)**.** If $G = (V_G, E_G)$ is a graph with $V_G = \{v_1, v_2, \ldots, v_n\}$ and w_G is a weight function associated with it, then one defines the *Laplacian matrix* of G, denoted as $L(G)$, in the following way:

$$L(G)_{ij} = \begin{cases} -w_G(v_i, v_j) & \text{if } i \neq j; \\ \sum_{k=1, k\neq i}^{n} w_G(v_i, v_k) & \text{if } j = i. \end{cases}$$

It is easy to see from Definition 1.2.12 that, if L is the Laplacian matrix of some graph, then

$$L\mathbf{1} = \mathbf{0}, \tag{1.2}$$

where $\mathbf{1}$ and $\mathbf{0}$ are the vectors whose entries are all equal to one and zero, respectively. It is easy to show that the converse is also true.

Proposition 1.2.1. *Every symmetric matrix L with non-positive off-diagonal elements satisfying*

$$L\mathbf{1} = \mathbf{0}$$

is a Laplacian matrix of some graph.

Proof. Let the matrix L be of size n. We can find the graph, $G = (V_G, E_G)$, which corresponds to L by letting $V_G = \{v_1, v_2, \ldots, v_n\}$ and associating a weight function, w_G, with G by letting

$$w_G(v_i, v_j) := -L_{ij}, \quad 1 \leq i < j \leq n.$$

Then it is easy to check that L is indeed the Laplacian matrix of the graph $G = \{V_G, E_G\}$ with weight function w_G. □

Proposition 1.2.2 (**Basic property of Laplacian matrices**). *Let $L \in \mathbb{R}^{n\times n}$ be a Laplacian matrix and $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ be some vectors. Then*

$$\mathbf{x}^T L\mathbf{y} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} L_{ij}(x_i - x_j)(y_i - y_j). \tag{1.3}$$

In particular, L is a positive semi-definite matrix.

Proof. From (1.2) we have that

$$L_{ii} = -\sum_{j=1, j\neq i}^{n} L_{ij}.$$

Therefore, if we let $S := \mathbf{x}^T L\mathbf{y}$, then

$$\begin{aligned} S &= \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}x_i y_j = \sum_{i=1}^{n} L_{ii}x_i y_i + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} L_{ij}x_i y_j \\ &= -\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} L_{ij}x_i y_i + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} L_{ij}x_i y_j = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}x_i(y_j - y_i). \end{aligned} \tag{1.4}$$

Hence, if we swap i and j in (1.4), we obtain

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ji}x_j(y_i - y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}x_j(y_i - y_j),$$

where we have used that L is symmetric. Thus,

$$2S = -\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}(x_i - x_j)(y_i - y_j)$$

and since $L_{ij}(x_i - x_j)(y_i - y_j) = L_{ji}(x_j - x_i)(y_j - y_i)$ and $L_{ii}(x_i - x_i)(y_i - y_i) = 0$ for

all $1 \leq i, j \leq n$, we obtain

$$S = \mathbf{x}^T L\mathbf{y} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}(x_i - x_j)(y_i - y_j) = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} L_{ij}(x_i - x_j)(y_i - y_j),$$

and in particular,

$$\mathbf{x}^T L\mathbf{x} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} L_{ij}(x_i - x_j)^2,$$

which implies that L is positive semi-definite matrix. □

Definition 1.2.13 (**Fiedler vector of Laplacian matrix**)**.** Let L be the *Laplacian matrix* of some connected graph and $0 = \lambda_1 < \lambda_2 < \lambda_3 \leq \cdots \leq \lambda_n$ be the eigenvalues of L. The unit vector $\mathbf{v}_2$, corresponding to the second smallest eigenvalue of L, λ_2, is called the *Fiedler vector* of L.

Sometimes *Fiedler vectors* of Laplacian matrices are also called *Fiedler vectors* of graphs, associated with the particular Laplacian matrix. The importance of the Fiedler vector in Graph Theory was first discovered in (Donath and Hoffman, 1973) and then in (Fiedler, 1973) and (Fiedler, 1975), from where it has received its name.

Definition 1.2.14 (**Normalised Laplacian matrix of a graph**)**.** Let $G = (V_G, E_G)$ be a graph with $V_G = \{v_1, v_2, \ldots, v_n\}$ and w_G be a weight function associated with it. Then the matrix $\hat{L}$, defined by

$$\hat{L}_{ij} := \begin{cases} -\frac{w_G(v_i, v_j)}{\sqrt{d_i}\sqrt{d_j}} & \text{if } i \neq j; \\ 1 & \text{if } i = j, \end{cases}$$

where

$$d_i = \sum_{k=1, k\neq i}^{n} w_G(v_i, v_k)$$

is called a *normalised Laplacian matrix* of G.

Definition 1.2.15 (**Normalised Fiedler vector of a normalised Laplacian matrix.**)**.** Similarly to the Fiedler vector of a Laplacian matrix, the *normalised Fiedler vector* is defined as the unit eigenvector, corresponding to the second smallest eigenvalue of the normalised Laplacian matrix.

Definition 1.2.16 (**Adjacency matrix of a graph**)**.** Let $G = (V_G, E_G)$ be a graph with $V_G = \{v_1, v_2, \ldots, v_n\}$ and w_G be a weight function associated with it. Then the

matrix A, defined by

$$A_{ij} := \begin{cases} w_G(v_i, v_j) & \text{if } i \neq j; \\ 0 & \text{if } i = j, \end{cases}$$

is called an *adjacency matrix* of G.

It is easy to see from Definitions 1.2.12 and 1.2.14 that the two Laplacian matrices, L and $\hat{L}$, and the adjacency matrix, A, satisfy the following relation:

$$\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I_n - D^{-\frac{1}{2}} A D^{-\frac{1}{2}},$$

where D is the diagonal matrix $diag(d_1, d_2, \ldots, d_n)$ and I_n is the identity matrix of size n.

1.2.3. Clustering graphs and the Fiedler vector

Problem 1.2.1. *Given a graph $G = (V_G, E_G)$ and a weight function, w, associated with it, the goal is, loosely speaking, to split the vertices of the graph into two disjoint sets, $V_G^{(1)}$ and $V_G^{(2)} = V_G \setminus V_G^{(1)}$, in such a way that the edges joining vertices from $V_G^{(1)}$ and $V_G^{(2)}$ are as few as possible, while the sets $V_G^{(1)}$ and $V_G^{(2)}$ are balanced.*

Mathematically, Problem 1.2.1 can be stated in many ways, since there could be many different notions of "balance" between $V_G^{(1)}$ and $V_G^{(2)}$. We shall give a brief review of a few of the most common ways of representing Problem 1.2.1 as an optimisation problem. We start with a definition.

Definition 1.2.17. Given a graph $G = (V_G, E_G)$, let A and $B = V_G \setminus A$ be two disjoint subsets of V_G. Then the quantity $cut(A, B)$ is the sum of the weights of the edges joining vertices from A and B, that is,

$$cut(A, B) := \sum_{u \in A, v \in B} w(u, v).$$

The first approach, developed by (Donath and Hoffman, 1973), (Fiedler, 1973) and (Fiedler, 1975), and later popularised by (Pothen et al., 1990), minimises $cut(A, B)$ with the requirement that the subgraphs A and B have the same number of vertices. In other words, we have the following optimisation problem:

$$\min\{cut(A, B) \mid A \subset V_G, B = V_G \setminus A \text{ and } |A| = |B|\}, \tag{1.5}$$

where $|A|$ denotes the number of elements in A. Obviously, when the order of the graph is an odd number, it is no longer possible to have $|A| = |B|$. Then the requirement

$|A| = |B|$ is replaced by $\big||A| - |B|\big| = 1$.

Following (Higham et al., 2007), we shall now illustrate the link between problem (1.5), the Laplacian matrix of G (c.f. Definition 1.2.12) and the Fiedler vector of G (c.f. Definition 1.2.13).

Let us assume that the graph G is of order n, that is, $|V_G| = n$, and $\mathbf{p} = (p_1, p_2, \ldots, p_n)^T$ is a vector whose entries are either -1 or 1, depending on whether a vertex belongs to the set A or B. Specifically, let

$$p_i := \begin{cases} 1 & \text{if } v_i \in A \\ -1 & \text{if } v_i \in B. \end{cases}$$

Then, from Proposition 1.2.2 we have

$$cut(A, B) = \frac{1}{4} \sum_{\{i,j|v_i \sim v_j\}} w(v_i, v_j)(p_i - p_j)^2 = \mathbf{p}^T L \mathbf{p},$$

where L is the Laplacian matrix associated with G. In terms of the vector $\mathbf{p}$, the constraints $|A| = |B|$ when n is even, or $\big||A| - |B|\big| = 1$ when n is odd, become

$$|\mathbf{p}^T \mathbf{1}| = \begin{cases} 0 & \text{when } n \text{ is even} \\ 1 & \text{when } n \text{ is odd.} \end{cases}$$

So, instead of problem (1.5) we may consider

$$\min\{\mathbf{p}^T L \mathbf{p} \mid p_i \in \{-1, 1\}, |\mathbf{p}^T \mathbf{1}| \leq \theta\}, \tag{1.6}$$

where the parameter $\theta \geq 0$ provides us with some flexibility regarding the balance between A and B. For example, (1.6) with $\theta = 1$ is equivalent to (1.5), since it only allows clusters A and B with $|A| = |B|$, when n is even, or $\big||A| - |B|\big| = 1$, when n is odd. Choosing θ larger on the other hand loosens the restriction of balance between A and B in (1.5), e.g. $\theta = n$ minimises $cut(A, B)$ over all possible choices of A and $B = V_G \setminus A$.

Unfortunately, problems (1.5) and (1.6) are NP-complete and therefore untractable, when the order of G, n, is very large. One way of approximating the solution of (1.6), is to *relax* the discrete optimisation problem by allowing the entries of the vector $\mathbf{p}$ to be real numbers, that is, $\mathbf{p} \in \mathbb{R}^n$. This turns (1.6) into a continuous optimisation problem:

$$\min\{\mathbf{p}^T L \mathbf{p} \mid \mathbf{p} \in \mathbb{R}^n, |\mathbf{p}^T \mathbf{1}| \leq \theta\}. \tag{1.7}$$

Since Laplacian matrices are positive semi-definite (c.f. Proposition 1.2.2), we still have

$\mathbf{p}^T L\mathbf{p} \geq 0$ for all vectors $\mathbf{p}$ satisfying the constraints in (1.7), as in the discrete case. However, unlike the discrete case, the solution to (1.7) would be some real vector. If there is inherent cluster structure in G, we normally expect the entries of the vector solving (1.7) to fall into distinct bands, so that a clustering emerges, resembling the way we cluster in the discrete case. But since the solution to (1.7) only "approximates" that of (1.6), there are cases when the clustering produced by (1.7) is very different from that produced by (1.6) (c.f. (Guattery and Miller, 1998)).

Firstly, we are interested in whether (1.7) has any nonzero solutions. Suppose the vector $\mathbf{p}_0 \neq 0$ solves (1.7) and $0 < \varepsilon < 1$. Then the vector $\varepsilon\mathbf{p}_0$ satisfies $|\varepsilon\mathbf{p}_0^T\mathbf{1}| < |\mathbf{p}_0^T\mathbf{1}| \leq \theta$ and provides a better "solution" to (1.7). This means that if we don't add a constraint, which normalises the size of $\mathbf{p}$, we shall always get $\mathbf{p} = \mathbf{0}$ as a solution to (1.7). In the discrete case we had $\|\mathbf{p}\|_2 = \sqrt{n}$, and therefore we can add that requirement to (1.7). Hence, (1.7) becomes

$$\min\{\mathbf{p}^T L\mathbf{p} \mid \mathbf{p} \in \mathbb{R}^n, |\mathbf{p}^T\mathbf{1}| \leq \theta, \|\mathbf{p}\|_2 = \sqrt{n}\},$$

which is equivalent to finding

$$\min\{\mathbf{p}^T L\mathbf{p} \mid \mathbf{p} \in \mathbb{R}^n, |\mathbf{p}^T\mathbf{1}| \leq \frac{\theta}{\sqrt{n}}, \|\mathbf{p}\|_2 = 1\}. \tag{1.8}$$

The solution to (1.8) is given by the following variation of the Rayleigh-Ritz Theorem (c.f. (Horn and Johnson, 1990), Theorem 4.2.2 and (Higham et al., 2007), Theorem 3.1):

Theorem 1.2.1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues ordered $\nu_1 < \nu_2 \leq \nu_3 \leq \cdots \leq \nu_n$ and corresponding mutually orthonormal eigenvectors $\mathbf{x}^{[1]}, \mathbf{x}^{[2]}, \ldots, \mathbf{x}^{[n]}$. Then, for fixed $0 \leq \alpha < 1$, the problem*

$$\min\{\mathbf{p}^T A\mathbf{p} \mid \mathbf{p} \in \mathbb{R}^n, |\mathbf{p}^T\mathbf{1}| \leq \alpha, \|\mathbf{p}\|_2 = 1\}$$

is solved by $\mathbf{p} = \alpha\mathbf{x}^{[1]} + \sqrt{1-\alpha^2}\mathbf{x}^{[2]}$.

Remark 1.2.4. *In order to apply Theorem 1.2.1 to some graph Laplacian matrix L, we have to make sure that the eigenvalue $\lambda_1 = 0$ is simple. The latter is equivalent to the requirement for the graph G to be* connected*, which we shall assume filfilled.*

Theorem 1.2.1, applied to problem (1.8), automatically yields that the minimum to the *relaxed* optimisation problem is achieved for

$$\mathbf{p} = \alpha\mathbf{1} + \sqrt{1-\alpha^2}\mathbf{v}_2, \tag{1.9}$$

where $\alpha = \frac{\theta}{\sqrt{n}}$ and $\mathbf{v}_2$ is the Fiedler vector of the graph G. As we mentioned earlier, we are looking for some structure in the elements of $\mathbf{p}$, in the sense of two or more bands of elements with significant gaps between them, resembling the discrete vector $\mathbf{p}$, which solves (1.6). This is achieved by firstly sorting the elements of $\mathbf{p}$ in an increasing order, and then looking for clusters in those elements. Each cluster of elements of $\mathbf{p}$ will correspond to a cluster of vertices in G, that is, if the entries p_i and p_j of $\mathbf{p}$ belong to different (the same) clusters, then the vertices v_i and v_j will also belong to different (the same) clusters. Therefore, in the solution (1.9) to the *relaxed* problem (1.8), the only source of discrimination between the different clusters of G is $\mathbf{v}_2$, the Fiedler vector of G, since the entries of the constant vector $\mathbf{1}$ only scale those of $\mathbf{p}$.

Above we have given a brief review of one of the commonly used techniques, that of *relaxation*, for obtaining information about graphs, using properties of the spectra of some matrix associated with them. As it was noted above, the main advantage of such an approach is that it usually modifies NP-hard or NP-complete problems on the graph into continuous problems, whose solution can often be found at little computational cost. The main goal of this thesis is the consideration of the sensitivity of the spectrum of some matrices associated with graphs, and not the analysis and derivation of spectral methods as such. However, in order to provide both motivation and background for the work carried out in this thesis, we list some of the approaches for analysing graphs via spectral properties of some matrix associated with them, also known as Spectral Graph Theory.

An extension to the idea of using the Fiedler vector for clustering graphs, which was briefly presented above, is given in (Higham et al., 2005), where the connection is made between the leading eigenvectors of a scaled version of the Adjacency matrix and the representation of the graph in a lower, usually two- or three-dimensional space. For example, the Fiedler vector in the approach above can be considered as one-dimensional representation, which preserves the cluster structure of the graph.

For any two disjoint subsets of vertices, A and B, such that $A \cup B = V_G$ (Ding et al., 2001) introduce the quantity

$$\mathrm{Mcut}(A, B) := \frac{\sum_{u \in A, v \in B} w(u, v)}{\sum_{u \in A, v \in A} w(u, v)} + \frac{\sum_{u \in A, v \in B} w(u, v)}{\sum_{u \in B, v \in B} w(u, v)},$$

and consider the problem of minimising $\mathrm{Mcut}(A, B)$ over all possible choices of A and $B = V_G \setminus A$. If we assume that minimum is achieved for $A = A^o$ and $B = B^o$, this approach will make the *cut* between A^o and B^o, that is, the quantity $\sum_{u \in A^o, v \in B^o} w(A^o, B^o)$, small, while "balancing" between the connectedness within A^o and B^o. Again, by *relaxing* the problem, (Ding et al., 2001) find that a "good approximation" to the optimal

solution is given by the Normalised Fiedler vector of the Normalised Laplacian of the graph (c.f. Definition 1.2.15).

(Shi and Malik, 2000) consider the quantity

$$\mathrm{Ncut}(A,B) := \frac{\sum_{u\in A, v\in B} w(u,v)}{\deg(A)} + \frac{\sum_{u\in A, w\in B}}{\deg(B)},$$

where $\deg(A) = \sum_{u\in A} d_u$ is the sum of the node degrees of subgraph A, and find that the Normalised Fiedler vector of the Normalised Laplacian of the graph provides a "good approximation" to the optimal solution.

(Newman, 2006) considers the following matrix

$$B_{ij} := w(v_i, v_j) - \frac{1}{2m} d_{v_i} d_{v_j}, \tag{1.10}$$

where $m = \frac{1}{2}\sum_{u\in V_G} d_u$ is the total number of edges in the network (which is not necessarily an integer). The entry B_{ij}, for $i \neq j$, represents the difference between the actual weight between vertices v_i and v_j, and the expected weight between those two vertices, if edges were placed at random (see (Newman, 2006) for more details). Then (Newman, 2006) considers the problem of maximising the quantity

$$Q := \frac{1}{4m}\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij} s_i s_j = \frac{1}{4m}\mathbf{s}^T B \mathbf{s}$$

over all vectors $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$, whose entries satisfy $s_1 \in \{-1, 1\}$. In other words, the goal is to separate G into two sets of vertices:

$$A = \{v_i \mid s_i = -1\} \qquad \text{and} \qquad B = \{v_j \mid s_j = 1\},$$

such that v_i and v_j are placed in the same set, if the weight of the edge between them, $w(v_i, v_j)$, is greater than what we would expect in a random graph, and otherwise v_i and v_j are placed in different sets. The quantity Q is called *modularity* of the graph G. Also, it is interesting to note that B, defined by (1.10), is symmetric and satisfies (1.2) but, generally speaking, is not a Laplacian matrix according to Definition 1.2.12, since not all of its off-diagonal entries have the same sign.

The problem

$$\max\{\mathbf{s}^T B \mathbf{s} \mid s_i \in \{-1, 1\}, 1 \leq i \leq n\}$$

is then *relaxed* in a way similar to that in (1.8) and the solution to the continuous problem is found to be $\mathbf{u}_1$, the eigenvector corresponding to the largest eigenvalue of B, if the latter is positive. An interesting feature of the *modularity* is that, if B turns

out to be a negative semi-definite matrix, then the solution to the *relaxed* problem is the vector $\mathbf{1}$ (since $B\mathbf{1} = \mathbf{0}$), which assigns all vertices to one cluster (or community) and this can be interpreted as lack of community structure in the graph (see (Newman, 2006) for further details).

Random walks on graphs are another example of the application of spectral properties of matrices to uncover important properties of graphs. Given a graph $G = (V_G, E_G)$ with some *consistent* weight function w on it (c.f. Definition 1.2.7), the *transition* matrix M, associated with G, is given by

$$M_{ij} = \begin{cases} \frac{w(v_i,v_j)}{d_{v_i}} & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

The (i, j)-th entry of the *stochastic* matrix M (c.f. (Horn and Johnson, 1990)) is the probability of jumping to vertex v_j, given we are at vertex v_i. It is well known (by the Perron-Frobenius theorem for non-negative matrices, c.f. (Horn and Johnson, 1990), Theorem 8.4.4) that the largest eigenvalue of M is real, simple and is equal to 1 if M is *irreducible*. Furthermore, the left eigenvector of M, say $\boldsymbol{\pi}$, corresponding to the eigenvalue 1 has nonnegative entries whose sum may be normalised to equal one. Suppose $\boldsymbol{\pi}$ is such a normalised eigenvector of M. Then

$$\boldsymbol{\pi} M = \boldsymbol{\pi}$$

and it can be shown that $\boldsymbol{\pi}$ represents the stationary distribution of the random walk on G, that is, the i-th component of $\boldsymbol{\pi}$ is the probability of being at the vertex v_i if we "randomly walk" on G for a long time. A technique similar to this, though a lot more sophisticated than what we have described, is used by Google for finding the ranks of the web pages (c.f. (Brin et al., 1998)).

1.3. Literature review.

To the best of our knowledge there is little literature, which considers the problem of sensitivity to random perturbations of symmetric matrices, leading to corresponding analysis for associated networks.

Random Matrix Theory (RMT) is a very active area of research, especially within the Quantum Physics community, with huge amount of literature, but little of this is relevant to our work. However, in §2 we have used a famous result by (Tracy and Widom, 1996), which provides a link between the solution to an initial value problem and the distribution of the largest eigenvalue of a matrix from the Gaussian Orthogonal

Ensemble (GOE).

Stewart in (Stewart, 1990) has given a detailed probabilistic approach to general Matrix Perturbation Theory. As an application of that, he considers the problem of sensitivity of the spectra of matrices to stochastic perturbations by *cross-correlated* matrices. In doing this he uses first-order perturbation expansion. However, as we show in §4, this approach can have some limitations when applied to the problems we consider.

Karrer et al. in (Newman et al., 2008) consider the robustness of modularity and community structure of networks subject to random changes of the positions of their edges, so that the number of vertices and edges in the network is preserved. They construct a measure on the difference between different community assignments by using *variation of information*. Based on that measure, they test the robustness of the community assignment of a given network by simulations, that is, by perturbing the network many times. Although very sophisticated and being able to detect subtle differences in community assignments, their information-theoretic distance metric appears to be difficult to apply to measuring the sensitivity of network clustering, without simulating the perturbation. Thus, since some of the networks arising in applications can be very large, we consider the approach in (Newman et al., 2008) impractical for our purposes.

1.4. Features of this Thesis.

As far as we know, this thesis is a first attempt at rigorous perturbation theory for the symmetric eigenvalue problem with stochastic perturbation. We have combined Linear Algebra with Probability Theory and recent results from the Spectral Theory of GOE matrices, to obtain bounds on the sensitivity to stochastic perturbation of both, eigenvalues and eigenvectors, of deterministic symmetric matrices. Most of these results are stated in §3 and §4, and are for scaled versions of GOE matrices, but they extend to more general classes of random matrices.

The results we obtain here are an attempt to provide an analytic tool, which substitutes simulations, for measuring the sensitivity of the clustering of networks, to random noise which can be modelled by certain types of random symmetric matrices. The advantages are that accuracy obtained by using theoretical (or analytical) results is very close to that of simulations, while the latter are computationally very costly. In fact, the results in §3 and §4 have the potential of even being tabulated for a prescribed range of dimensions, which would make the computational cost of obtaining them, or interpolating in order to obtain them, negligible, compared to that of simulations. This

may also be considered as a first step towards creating a competitive measure on the robustness of spectral clustering, or modularity and community structure of networks (c.f. (Rand, 1971), (Newman et al., 2008) and (Newman, 2006)), without the use of simulations.

In §5 we consider networks of coupled oscillators. This represents another aspect of network theory, which combines the dynamics of the vertices with the spectral properties of the network. An interesting consequence of this is that large and complex networks of oscillators can be reduced to smaller networks, consisting of only a few oscillators, from which the dynamics of the rest of the oscillators in the network can be deduced. Such dynamical structures are called "master-slave" systems. We also give an algorithm by which such "master-slave" can be detected in networks.

In §6 we make a first attempt at comparing rigorously spectral clustering by Laplacian and normalised Laplacian matrices. This is done on Path graphs and on products of Path graphs. In the analysis we use Homotopy between the two clusterings.

1.5. Plan of this Thesis.

In §2.2 we revise known theory for approximating the distribution of the largest eigenvalue of a GOE matrix. In §2.3 we state a conjecture, from which we derive an asymptotic relation between the distributions of the largest eigenvalue of a GOE matrix and its 2-norm. From this we derive a numerical procedure, which helps us to approximate the distribution of the 2-norm of a GOE matrix, using theory for the distribution of its largest eigenvalue. Numerical experiments, used to justify the conjecture, agree well with the theory. In §2.4 we state an extension to the conjecture in §2.3, which covers a broader class of random symmetric matrices. Next, we state a third conjecture, which implies a relation between the distributions of the largest eigenvalue of a random (symmetric) Laplacian matrix and its 2-norm. Both conjectures in §2.4 are tested numerically and results indicate good agreement with the theory. In §2.6 we further extend the class of matrices, whose 2-norm's distribution can be approximated numerically, by considering matrices which differ from GOE matrices only in their diagonal elements. Again, numerical experiments support the theory.

In §3 we consider the problem of finding the maximum possible magnitude of perturbation, such that the eigenvalues of a symmetric matrix, perturbed by a random matrix, do not swap with a given probability. In §3.2 we revise known results in Linear Algebra. In §3.3 we combine Markov's inequality with Bauer-Fike's Theorem to provide a bound on the size of the allowable perturbation. In §3.4 we improve the bound found in §3.3, by using the approximation of the distribution of the 2-norm of SGOE

matrices, given in §2. In §3.5 we combine a simple residual theorem and the Bauer-Fike Theorem to further improve the bound on size of the allowable perturbation. Further, we suggest a way of finding this bound numerically, by assuming that the 2-norm and the norm of a row, or a column, of SGOE matrix become "less dependent", as the size of the matrix increases. Numerical tests in §3.9 support the assumption. In §3.6 we provide a comparison between the bounds found in §3.4 and §3.5, and show that the latter is larger asymptotically. In §3.7 we state a stochastic analogue of the deterministic version of the $\sin\psi$ theorem, which provides both upper and lower bounds on the angle between an eigenvector and its perturbed counterpart. In §3.8 we extend the result in §3.5 to the stochastic perturbation of singular values of rectangular matrices. Finally, in §3.9 we test and compare the bounds found in §3.3, §3.4 and §3.5.

In §4 we address the inverse of the problem considered in §3. Namely, given the size of perturbation, we provide upper bounds on the probability of a swap for simple eigenvalues. We consider two approaches. Firstly, in §4.2 we use first-order Perturbation Theory to derive an upper bound on the probability of a swap. The theory in that section is tested numerically in §4.4 and, as a conclusion, we show the limitations of this approach. Secondly, in §4.3 we consider an approach, similar to that in §3.5 in that it uses the same deterministic results as a basis. However, at the end of §4.3 we argue that the upper bound on the probability of a swap, provided by this second approach, is very crude and thus impractical. Numerical experiments, which we do not provide, confirm the limited practical application of this bound.

In §5 we consider entrainment of networks of coupled oscillators. By combining spectral properties of the matrix, associated with the network of oscillators, and their dynamics, we show that the groups of oscillators which entrain, when the size of the strength parameter is large, can be determined by considering the entries of the second largest eigenvector of the matrix, associated with the network. A numerical experiment on a range-dependant network (c.f. (Grindrod, 2002)) of oscillators supports that theory. Further, in §5.4 we consider special case of coupled systems of oscillators. Namely, where the coupling is such that there are "master" and "slave" oscillators, with directed edges from the master set to the slave set, but no directed edges from the slaves back to the masters. In §5.4.1, using the Perron-Frobenius Theorem, we suggest an algorithm for uncovering "master-slave" structures in networks of coupled oscillators, by the zero entries in left and right eigenvectors of the adjacency matrix of the network. In §5.4.2 we investigate the implications of a "master-slave" structure to the entrainment between the oscillators in the network. This algorithm given in §5.4.1 is tested on a small example in §5.5. Results from the experiment agree with the theory.

In §6 we compare clusterings of Path graphs and products of Path graphs with respect to the Laplacian and the normalised Laplacian matrices, associated with the graphs. The theory in this section is an attempt to answer the question of whether there are, in general, any significant differences between the clusterings obtained by these two matrices. (Higham et al., 2007) suggest that clustering with respect to normalised Laplacian matrices tends to make the clustering "less susceptible to the influence of "poorly calibrated" vertices that have abnormally large or small weights" (quote from (Higham et al., 2007)). By using properties of symmetric tridiagonal matrices and properties of the Kronecker product of matrices, we conclude that there is no significant difference when products of Path graphs are clustered by the Laplacian and the normalised Laplacian matrices.

Chapter 2. Distribution of the norm of a Random Symmetric Matrix

2.1. Introduction.

In this chapter we give a way of approximating the distribution of the 2-norm of a random symmetric matrix from the Gaussian Orthogonal Ensemble (GOE). This is then used in the next chapter, where we extend standard deterministic matrix perturbation theorems for symmetric matrices to the case of stochastic perturbations.

The material in this chapter is based on the remarkable results of Tracy and Widom (c.f. (Tracy and Widom, 1994b) and (Tracy and Widom, 1996)), where the connection is made between asymptotic behaviour of the cumulative distribution function (c.d.f.) of the maximum eigenvalue of a GOE matrix and the solution to a Painlevé II initial value problem (see Definition A.1.6, where we define *cumulative distribution function*, abbreviated as *c.d.f.*). A MATLAB program is given by Edelman and Persson (c.f. (Edelman and Persson, 2005)) which computes the c.d.f. of the largest eigenvalue of a GOE matrix by solving the initial value problem and numerical experiments (given in §2.2) show that the numerical solution of the initial value problem gives a good agreement with the asymptotic behaviour for surprisingly low dimensional GOE matrices.

In this chapter we extend these results in the following directions. Firstly, we make a conjecture (see Conjecture 2.3.1) relating the distribution of the 2-norm of a GOE matrix to the distribution of its largest eigenvalue. Numerical results (given in §2.3) for matrices of both large and small dimension support this conjecture. Secondly, using Conjecture 2.3.1 and a minor extension of the MATLAB program of Edelman and Persson, we compute the asymptotic form of the c.d.f. and p.d.f. of the 2-norm of a GOE matrix (for a definition of *probability density function*, or *p.d.f.*, see Definition A.1.7). Numerical results again indicate good agreement with simulations. We discuss the considerable computation advantages of the approach using the solution of Painlevé II at the end of §2.3. Thirdly, we extend the class of matrices whose 2-norm's c.d.f.'s and p.d.f.'s can be calculated using appropriate extensions of Conjecture 2.3.1

and Program 2.3.1 (see §2.4). Fourthly, we further extend that class of matrices by considering matrices which differ from GOE matrices only in their diagonal elements.

Finally, we mention some notation. We denote asymptotic forms of the c.d.f. and p.d.f. obtained by solving Painlevé II by $G(s)$ and $g(s)$, respectively. After conversion to n dimensions by a scaling variable (see (2.5)) we denote the c.d.f. and p.d.f. of the 2-norm of an n-dimensional GOE matrix by $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$. Lastly, we denote the c.d.f. and p.d.f. of the 2-norm of an n-dimensional GOE matrix, obtained by simulation (i.e. by repeated experiments with a large number of samples of GOE matrices) by $G_n^{(S)}(t)$ and $g_n^{(S)}(t)$, respectively. In our experiments we test the reliability of our theories by comparing $G_n^{(P)}(t)$ with $G_n^{(S)}(t)$ and $g_n^{(P)}(t)$ with $g_n^{(S)}(t)$.

The definitions and results in Probability Theory, which we use in this chapter, are given in the Appendix, §A.1.

2.2. Definitions and background.

We start with a definition.

Definition 2.2.1 (GOE matrix). We say that the symmetric matrix $B \in \mathbb{R}^{n\times n}$ belongs to GOE (Gaussian Orthogonal Ensemble), or that B is GOE matrix, if its entries are independent random variables[1] satisfying

$$\begin{array}{|ll} B_{ij} \in \mathcal{N}(0, \frac{1}{2}) & \text{for } 1 \le i < j \le n; \\ B_{ij} = B_{ji} & \text{for } 1 \le i < j \le n; \\ B_{ii} \in \mathcal{N}(0, 1) & \text{for } 1 \le i \le n. \end{array}$$

Remark 2.2.1. *In Definition 2.2.1 we have used $\mathcal{N}(\mu, \sigma^2)$ to denote a normally distributed random variable with mean (or expectation) μ and variance σ^2. See Definition A.1.11, where we define* normal distribution.

Let B be GOE matrix and

$$\beta_1 \le \cdots \le \beta_n$$

be its eigenvalues. Since B is GOE matrix, then (by Definition 2.2.1) the matrix $-B$ is also GOE, and hence the distribution of $-\beta_1$ must be the same as that of β_n. By definition $\|B\|_2 = \max\{|\beta_1|, |\beta_n|\} \ge \beta_n$. In fact, we shall see in §2.3 that $\|B\|_2$ also behaves like $\sqrt{2n}$ as $n \to \infty$. In these settings we also prove (in §2.6) that the diagonal entries of B play no role when we calculate $\lim_{n\to\infty} \sqrt{\frac{2}{n}}\|B\|_2$. The last result broadens the class of matrices, B, by which we could perturb a given deterministic symmetric A, and apply the stochastic version of the Bauer-Fike Theorem (see §3).

[1] We define the term *independent random variables* in Definition A.1.5

"Gaussian Orthogonal Ensemble" comes from a very important property of this class of matrices, namely their invariance w.r.t. orthogonal matrices. That is, if V is a deterministic orthogonal matrix and B is GOE matrix, then the matrix VBV^T is also GOE matrix. Although most authors in the area mention this property as central for the class of GOE matrices, and state that it is somehow "obvious", we haven't found any proofs of this result in the literature. This is why we present a proof of this invariance (see Theorem 2.5.1) and give some of its implications in the next chapter.

It is a standard result (c.f. §1.3 in (Forrester, 1993)) that the probability that the eigenvalues of the matrix B lie in an infinitesimal interval about the points $x_1, \ldots, x_n$ is given by

$$\mathbb{P}_n(x_1, \ldots, x_n) = C_n \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \prod_{j<k} |x_j - x_k|^2 dx_1 \cdots dx_n, \tag{2.1}$$

where C_n is a normalisation constant, such that

$$C_n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \prod_{j<k} |x_j - x_k|^2 dx_1 \cdots dx_n = 1.$$

Therefore, given $\mathbb{P}_n(x_1, \ldots, x_n)$, the cumulative distribution function (c.d.f.) of the largest eigenvalue of B, denoted here as $F_n(t)$, satisfies

$$F_n(t) := \mathbb{P}[\max\{\beta_1, \ldots, \beta_n\} < t] = \mathbb{P}\left[\bigcap_{i=1}^{n} \{\beta_i < t\}\right].$$

Hence

$$F_n(t) = \mathbb{P}[\beta_n < t] = \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} \mathbb{P}_n(x_1, \ldots, x_n) dx_1 \cdots dx_n. \tag{2.2}$$

Formula (2.2) is a nice way to represent the c.d.f. of β_n theoretically. However, it is impractical if one wants to calculate $F_n(t)$ for large values of n, since it involves multidimensional integration. In (Tracy and Widom, 1996) the authors give a way of approximating $F_n(t)$ by solving a Painlevé II ODE. The following paragraph sets up the scene, before we summarise their main result in (2.9) below.

The following result (c.f. (Bai and Yin, 1988))

$$F_n(\sqrt{2n} + x) \to \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} \qquad \text{as } n \to \infty \tag{2.3}$$

shows that the distribution of β_n concentrates around $\sqrt{2n}$. This is the reason for

introducing the *edge scaling variable*, s, defined in (2.5) below (see also (Bowick and Brezin, 1991), (Forrester, 1993) and (Tracy and Widom, 1994a)). If we define

$$\zeta_n := n^{1/6}\sqrt{2}(\beta_n - \sqrt{2n}), \tag{2.4}$$

(Tracy and Widom, 2000b) show that the sequence of random variables, $\{\zeta_n\}_{n\in\mathbb{N}}$, converges in distribution to a random variable, say ζ, which has a stationary distribution. In other words, since the distribution of β_n concentrates around the point $\sqrt{2n}$ as n increases, which can be seen from equation (2.3), we normalise it and instead consider ζ_n. This corresponds to replacing the variable t by a variable s, where the relation between the two is given by

$$t =: \frac{1}{\sqrt{2}}\left(2\sqrt{n} + \frac{s}{n^{1/6}}\right). \tag{2.5}$$

In (2.5) t represents the old scaling, expressed in terms of the new one. We further define (c.f. (Tracy and Widom, 1994a) and (Tracy and Widom, 1994b))

$$F(s) := F_\zeta(s) = \lim_{n\to\infty} F_n(t), \tag{2.6}$$

where $F(s)$ can be found by solving a Painlevé II ODE (described below), with s related to t via (2.5). The solution $F(s)$ is then used as an approximation of $F_n(t)$. This is clarified in the next paragraph. Numerical results in Experiment 2.2.1 show that this approximation is good even for small values of n.

Let q be the solution of the Painlevé II equation

$$q'' = sq + 2q^3 \tag{2.7}$$

with the asymptotic condition

$$q(s) \sim \mathrm{Ai}(s), \qquad \text{as } s \to \infty, \tag{2.8}$$

where Ai is the Airy function. We recall that the Airy function $\mathrm{Ai}(s)$ is one of the two linearly independent solutions to the ODE

$$\frac{d^2y}{ds^2} - sy = 0$$

and for real values of s $\mathrm{Ai}(s)$ is given by

$$\mathrm{Ai}(s) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{t^3}{3} + st\right) dt.$$

The Painlevé II equation (2.7) is a nonlinear Airy equation, defined on $-\infty < s < \infty$. In many cases of interest one asymptotic condition is given either as $s \to -\infty$ or as $s \to +\infty$, and the task is to determine the asymptotic behaviour at the other end. In the Tracy-Widom work the asymptotic condition is (2.8), but we are interested in the form of the solution, since this allows us to compute the cumulative distribution function $F(s)$ over $-\infty < s < \infty$. For the numerical solution, the condition (2.8) is converted to conditions on q and q' for a large value of s which, together with (2.7), provides an initial value problem to be solved backwards in s.

The main result is that

$$F(s) = \exp\left(-\frac{1}{2}\left(\int_s^\infty (x-s)q(x)^2 dx + \int_s^\infty q(x)dx\right)\right) \tag{2.9}$$

(cf. (Tracy and Widom, 2000a)). Once we obtain $F(s)$, we can easily return to $F_n(t)$ by using the relation between t and s given in (2.5). Therefore, given a t, we approximate $F_n(t)$ by taking

$$F_n^{(P)}(t) := F(s). \tag{2.10}$$

The discussion so far can be summarised in the following theorem:

Theorem 2.2.1 (c.f. (Tracy and Widom, 2000a))**.** *Let β_n be the largest eigenvalue of the matrix B, where B is from the Gaussian Orthogonal Ensemble (GOE). The normalised eigenvalue, ζ_n, is defined as*

$$\zeta_n := n^{1/6}\sqrt{2}(\beta_n - \sqrt{2n}).$$

Then, as $n \to \infty$,

$$\zeta_n \xrightarrow{\mathcal{D}} F(s),$$

where $F(s)$ is given by (2.9) *and $q(s)$ is the solution to the Painlevé II ODE* (2.7) *with boundary conditions* (2.8)*.*

In this paragraph we give an insight of where the relation between Painlevé II equations and the distribution of the largest eigenvalue of random GOE matrices comes from. It can be shown that the joint density function, $\mathbb{P}_n(x_1, x_2, \ldots, x_n)$, satisfies (c.f. (Mehta, 1991), (Tracy and Widom, 1994a))

$$\mathbb{P}_n(x_1, x_2, \ldots, x_n) = \frac{1}{n!}\det(K_n(x_i, x_j))_{i,j=1,\ldots,n},$$

where

$$K_n(x, y) = \sum_{k=0}^{n-1} \phi_k(x)\phi_k(y)$$

and the sequence $\{\phi_k(x)\}_{k\in\mathbb{N}}$ is obtained by orthonormalising the sequence

$$\left\{x^k \exp(-x^2)\right\}_{k\in\mathbb{N}}$$

over $(-\infty,\infty)$ (c.f. (Tracy and Widom, 1994a)). Because of its representation in terms of orthogonal polynomials, the Fredholm determinant of $K_n(x,y)$, in the limit when $n\to\infty$, can be given as a solution to a completely integrable system of PDEs, found by Jimbo, Miwa, Mori and Sato. Hence, the distribution of the largest eigenvalue of a GOE matrix can be given as a solution to a Painlevé II ODE. The link between orthogonal polynomials and completely integrable systems can be explained as follows: Because of the recurrence relation between orthogonal polynomials, they are integrable in the sense that they have a Lax pair formulation (c.f. (Fokas et al., 1992)). Thus, one expects their characteristics to be expressed by integrable equations, and the prototypical integrable ODEs are the Painlevé equations.

In (Edelman and Persson, 2005) the authors give a numerical procedure, implemented in MATLAB, which solves (2.7) with initial condition (2.8), and finds $F(s)$ over a discrete set of values of s. In fact, (2.8) is differentiated to give

$$q'(s) \sim \mathrm{Ai}'(s),$$

so we obtain two conditions for large s. The ODE (2.7), together with the two conditions derived from (2.8), is written as a system of ODEs and is solved backwards for a large value of s. Following (Edelman and Persson, 2005), we obtain the following system of ODEs

$$\frac{d}{ds}\begin{pmatrix} q \\ q' \\ I \\ I' \\ J \end{pmatrix} = \begin{pmatrix} q' \\ sq+2q^3 \\ I' \\ q^2 \\ -q \end{pmatrix} \tag{2.11}$$

with initial conditions

$$\begin{pmatrix} q(s_0) \\ q'(s_0) \\ I(s_0) \\ I'(s_0) \\ J(s_0) \end{pmatrix} = \begin{pmatrix} \mathrm{Ai}(s_0) \\ \mathrm{Ai}'(s_0) \\ \int_{s_0}^{\infty}(x-s_0)\mathrm{Ai}(x)^2dx \\ \mathrm{Ai}(s_0)^2 \\ \int_{s_0}^{\infty}\mathrm{Ai}(x)dx \end{pmatrix}, \tag{2.12}$$

where

$$I(s) := \int_s^\infty (x-s)q(x)^2 dx \qquad \text{and} \qquad J(s) := \int_s^\infty q(x)dx,$$

and s_0 is chosen large enough. In (Edelman and Persson, 2005) the authors have chosen $s_0 = 5$ and they solve backwards for s in the range $-8 \leq s \leq 5$. After introducing the functions $I(s)$ and $J(s)$ as part of the system of ODEs, the desired output, $F(s)$, is recovered using

$$F(s) = \exp\left(-\frac{1}{2}(I(s) + J(s))\right). \tag{2.13}$$

The main part of the MATLAB program, which solves (2.11) with initial conditions (2.12), is presented as Program 2.2.1 below. It has been copied from (Edelman and Persson, 2005).

Program 2.2.1.

```
deq=inline('[y(2);s*y(1)+2*y(1)^3;y(4);y(1)^2;-y(1)]','s','y');

s0=5;
sn=-8;
sspan=linspace(s0,sn,1000);

y0=[airy(s0); airy(1,s0); ...
    quadl(inline('(x-s0).*airy(x).^2','x','s0'),s0,20,1e-25,0,s0); ...
    airy(s0)^2; quadl(inline('airy(x)'), s0, 20, 1e-18)];

opts=odeset('reltol', 1e-13, 'abstol', 1e-15);
[s,y]=ode45(deq, sspan, y0, opts);

F = sqrt(exp(-y(:,3)-y(:,5)));
```

Experiment 2.2.1. *In this experiment we test the MATLAB algorithm suggested in (Edelman and Persson, 2005), presented in Program 2.2.1 above, for* $n = 5, 10, 20, 100$ *and* 500*, and compare its results against simulations of* β_n*. More precisely,* $F(s)$ *is obtained by solving the systems of ODEs* (2.11) *with initial conditions* (2.12) *numerically, using the MATLAB code in Program 2.2.1. Then the approximation of* $F_n(t)$*,* $F_n^{(P)}(t)$*, is obtained from* $F(s)$ *by using* (2.10)*. Further,* $F_n^{(P)}(t)$ *is compared with the c.d.f. of* β_n *obtained by simulation, here denoted as* $F_n^{(S)}(t)$*. In order to obtain* $F_n^{(S)}(t)$ *we simulate* 10 000 *samples of the GOE matrix* B *and for each of them we find* β_n *numerically. Then, using the built-in MATLAB function* `ecdf`*, we find the c.d.f. of the*

	$n = 10$	$n = 20$	$n = 100$	$n = 200$	$n = 500$
$\max_t \lvert F_n^{(S)}(t) - F_n^{(P)}(t)\rvert$	0.1183	0.0905	0.0718	0.0362	0.0302

Table 2.1: *The values of* $\max_t \left|F_n^{(S)}(t) - F_n^{(P)}(t)\right|$ *for* $n = 5, 10, 20, 100$ *and* 500.

vector, whose entries are the 10 000 *samples of* β_n. *The plots of* $F_n^{(P)}(t)$ *and* $F_n^{(S)}(t)$ *are presented in Figures 2-1, 2-2, 2-3, 2-4 and 2-5, where we also give the value of* $\max_t |F_t^{(P)}(t) - F_n^{(S)}(t)|$ *for each* n.

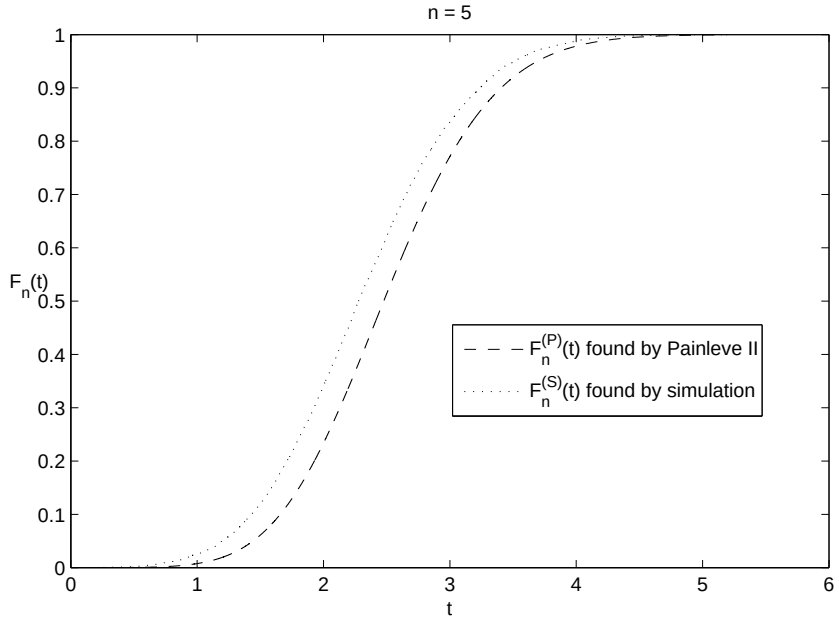

Figure 2-1: *Comparison between the c.d.f. of* β_n, $F_n(t)$, *found by simulation* ($F_n^{(S)}(t)$) *and as a solution to the Painlevé II ODE* ($F_n^{(P)}(t)$). *Here* $n = 5$ *and* $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)| = 0.1183$.

Results and Discussion. *In Table 2.1 we give the differences* $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)|$ *for different values of* n, $n = 5, 10, 20, 100$ *and* 500, *in order to check the rate of the "uniform" convergence of* $F_n^{(S)}(t)$ *to* $F_n^{(P)}(t)$. *By examining ratios of the form* $\frac{\max_t |F_{n_1}^{(S)}(t) - F_{n_1}^{(P)}(t)|}{\max_t |F_{n_2}^{(S)}(t) - F_{n_2}^{(P)}(t)|}$, *where* n_1 *and* n_2 *are different values of* n, *and comparing those ratios with* $\sqrt{\frac{n_1}{n_2}}$, *we can cnclude that up to* $n = 100$ *the rate of convergence behaves almost like* $\frac{1}{\sqrt{n}}$. *For example,*

$$\frac{\max_t |F_5^{(S)}(t) - F_5^{(P)}(t)|}{\max_t |F_{10}^{(S)}(t) - F_{10}^{(P)}(t)|} = \frac{0.1183}{0.0905} = 1.3072, \quad \frac{\max_t |F_{10}^{(S)}(t) - F_{10}^{(P)}(t)|}{\max_t |F_{20}^{(S)}(t) - F_{20}^{(P)}(t)|} = \frac{0.0905}{0.0718} = 1.2604$$

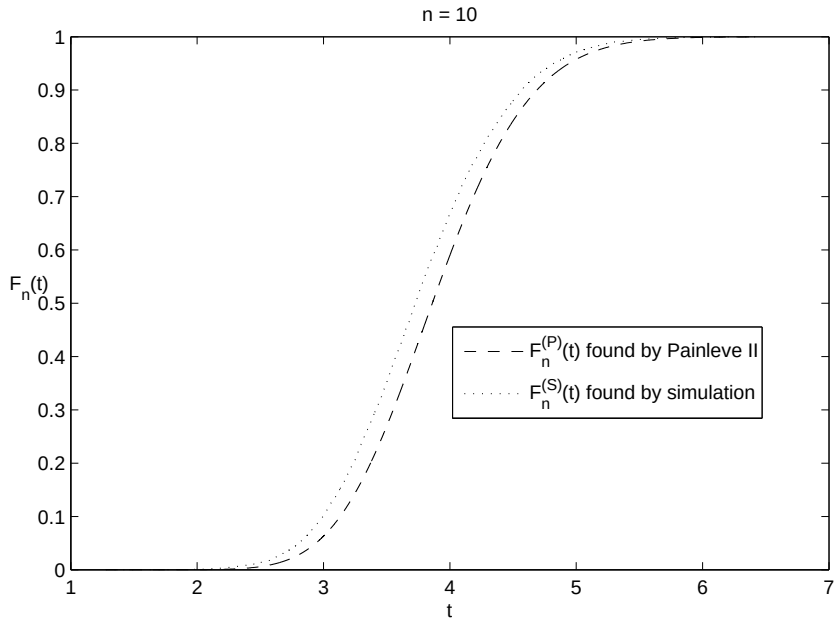

Figure 2-2: *Comparison between the c.d.f. of β_n, $F_n(t)$, found by simulation ($F_n^{(S)}(t)$) and as a solution to the Painlevé II ODE ($F_n^{(P)}(t)$). Here $n = 10$ and $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)| = 0.0905$.*

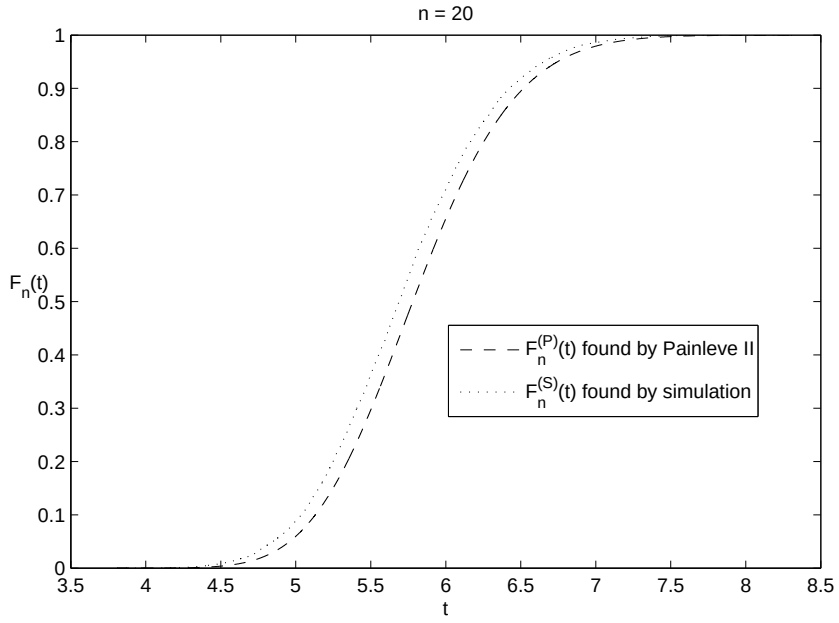

Figure 2-3: *Comparison between the c.d.f. of β_n, $F_n(t)$, found by simulation ($F_n^{(S)}(t)$) and as a solution to the Painlevé II ODE ($F_n^{(P)}(t)$). Here $n = 20$ and $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)| = 0.0718$.*

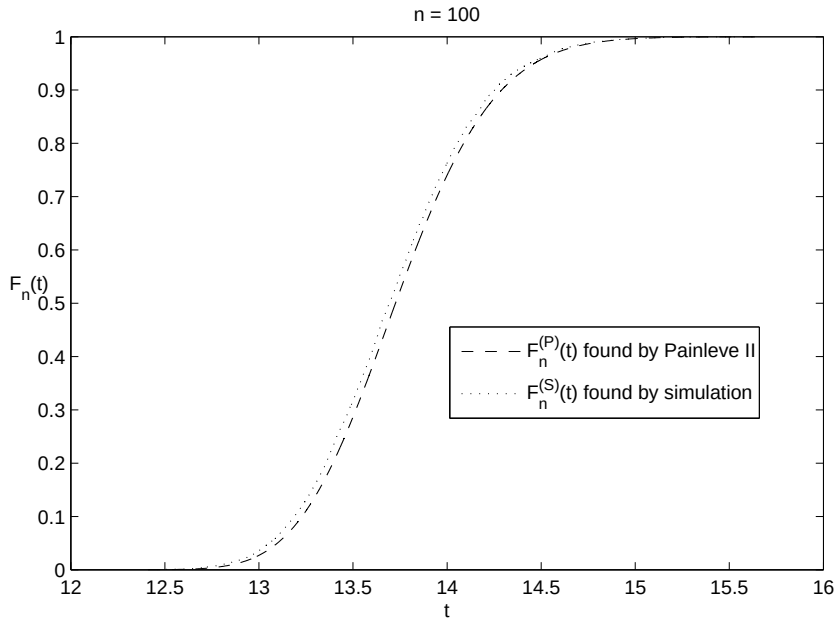

Figure 2-4: *Comparison between the c.d.f. of* β_n*,* $F_n(t)$*, found by simulation* ($F_n^{(S)}(t)$) *and as a solution to the Painlevé II ODE* ($F_n^{(P)}(t)$)*. Here* $n = 100$ *and* $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)| = 0.0362$.

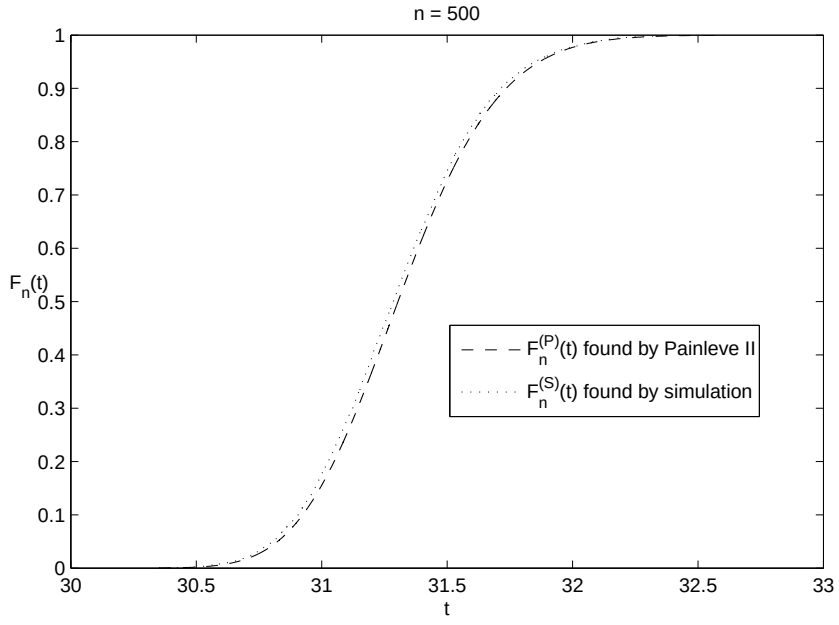

Figure 2-5: *Comparison between the c.d.f. of* β_n*,* $F_n(t)$*, found by simulation* ($F_n^{(S)}(t)$) *and as a solution to the Painlevé II ODE* ($F_n^{(P)}(t)$)*. Here* $n = 500$ *and* $\max_t |F_n^{(S)}(t) - F_n^{(P)}(t)| = 0.0302$.

and both, 1.3072 *and* 1.2604*, are close to* $\sqrt{2} \approx 1.4142$*. Also,*

$$\frac{\max_t |F_{20}^{(S)}(t) - F_{20}^{(P)}(t)|}{\max_t |F_{100}^{(S)}(t) - F_{100}^{(P)}(t)|} = \frac{0.0718}{0.0362} = 1.9834,$$

which is close to $\sqrt{5} \approx 2.2361$*. However, this rate of convergence doesn't seem to hold once the dimension,* n*, becomes* 500*, because*

$$\frac{\max_t |F_{100}^{(S)}(t) - F_{100}^{(P)}(t)|}{\max_t |F_{500}^{(S)}(t) - F_{500}^{(P)}(t)|} = \frac{0.0362}{0.0302} = 1.1987,$$

which is much smaller than $\sqrt{5}$*. This "phenomenon" is again observed in Table 2.2, in the case of GOE matrices. As we discuss later, the way to resolve this problem seems to be by increasing the number of simulations.*

2.3. The distribution of the 2-norm of GOE matrices.

We now concentrate on our main goal in this chapter, namely finding the distribution function of $\|B\|_2$,

$$G_n(t) := \mathbb{P}[\|B\|_2 < t]. \tag{2.14}$$

The link between the largest and smallest eigenvalues of B, β_n and β_1 respectively, and $\|B\|_2$ is given by the definition $\|B\|_2 := \max\{|\beta_1|, |\beta_n|\}$. We now make two observations about the distributions of $|\beta_1|$ and $|\beta_n|$.

Firstly, we recall that if B is a GOE matrix, then so is $-B$. This symmetry could also be seen from $\mathbb{P}_n(x_1, \ldots, x_n) = \mathbb{P}_n(-x_1, \ldots, -x_n)$ (see (2.1)). Thus, we may conclude that

$$\mathbb{P}[\beta_1 > -t] = \mathbb{P}[\beta_n < t] \qquad \text{and therefore} \qquad \mathbb{P}[|\beta_1| < t] = \mathbb{P}[|\beta_n| < t] \tag{2.15}$$

for all $t \in \mathbb{R}$. Secondly, from the limit in (2.3) we have

$$\mathbb{P}[\beta_n < 0] = F_n(0) \to 0, \quad \text{as } n \to \infty. \tag{2.16}$$

One could see from Figures 2-1, 2-2, 2-3, 2-4 and 2-5 that the convergence in (2.16) can effectively be taken as

$$\mathbb{P}[\beta_n < 0] = 0 \tag{2.17}$$

for values of n as small as 5. Therefore it is reasonable to assume $|\beta_n| = \beta_n$. Thus, using symmetry, $|\beta_1| = -\beta_1$ and hence, $-\beta_1$ should have the same distribution as β_n.

Therefore, if we define $G_n(t) := \mathbb{P}[\|B\|_2 < t]$, under the assumption (2.17) we obtain

$$G_n(t) = \mathbb{P}[\max\{|\beta_1|, |\beta_n|\} < t] = \mathbb{P}[|\beta_1| < t, \beta_n < t], \tag{2.18}$$

where $|\beta_1|$ and β_n have the same distribution (as a consequence of (2.17)).

In general, one cannot calculate $G_n(t)$, given in (2.18), without further knowledge about how $|\beta_1|$ and β_n depend on each other, even though they have the same distribution. However, our intuition tells us that, as n becomes larger, the dependance between the event $\{|\beta_1| < t\}$ and the event $\{\beta_n < t\}$ should "decrease". This is the statement of the following conjecture.

Conjecture 2.3.1. *If $B \in \mathbb{R}^{n\times n}$ is a GOE matrix and β_1 and β_n are its minimal and maximal eigenvalues respectively, then for all $t \in \mathbb{R}$ we have*

$$\mathbb{P}[|\beta_1| < t, |\beta_n| < t] \overset{n}{=} \mathbb{P}[|\beta_n| < t]^2, \tag{2.19}$$

where the notation "$\overset{n}{=}$" in (2.19) *means that*

$$\lim_{n\to\infty} \left|\mathbb{P}[|\beta_1| < t, |\beta_n| < t] - \mathbb{P}[|\beta_n| < t]^2\right| = 0.$$

In the notation in this chapter the statement of Conjecture 2.3.1 is equivalent to

$$G_n(t) \overset{n}{=} F_n(t)^2, \tag{2.20}$$

for all $t \in \mathbb{R}$.

Justification. (Based on experiments.) We test (2.19), or equivalently (2.20), by simulation. For a given n we simulate 10 000 samples of the $n \times n$ GOE matrix B. For each of these samples of B we calculate and store β_1 and β_n, its smallest and largest eigenvalues. We form two vectors, each of which contains 10 000 entries. The entries of the first vector are the samples of β_n and the entries of the second one are the samples of $\|B\|_2 = \max\{|\beta_1|, |\beta_n|\}$. From these two vectors, using the MATLAB built-in function `ecdf`, we calculate $F_n^{(S)}(t)$ and $G_n^{(S)}(t)$, the empirical c.d.f.'s of β_n and $\|B\|_2$, respectively. We then compare $F_n^{(S)}(t)^2$ and $G_n^{(S)}(t)$ by plotting them (see Figures 2-6, 2-7, 2-8, 2-9 and 2-10). The MATLAB program which we use for this simulation is given in the Appendix, Program A.2.1 there.

From Figures 2-6, 2-7, 2-8, 2-9 and 2-10 we can see that even for $n = 5$ the difference between $F_n^{(S)}(t)^2$ and $G_n^{(S)}(t)$ is very little and for $n \geq 20$ they visually coincide. This is a significant evidence that Conjecture 2.3.1, or equivalently (2.20), must be true. □

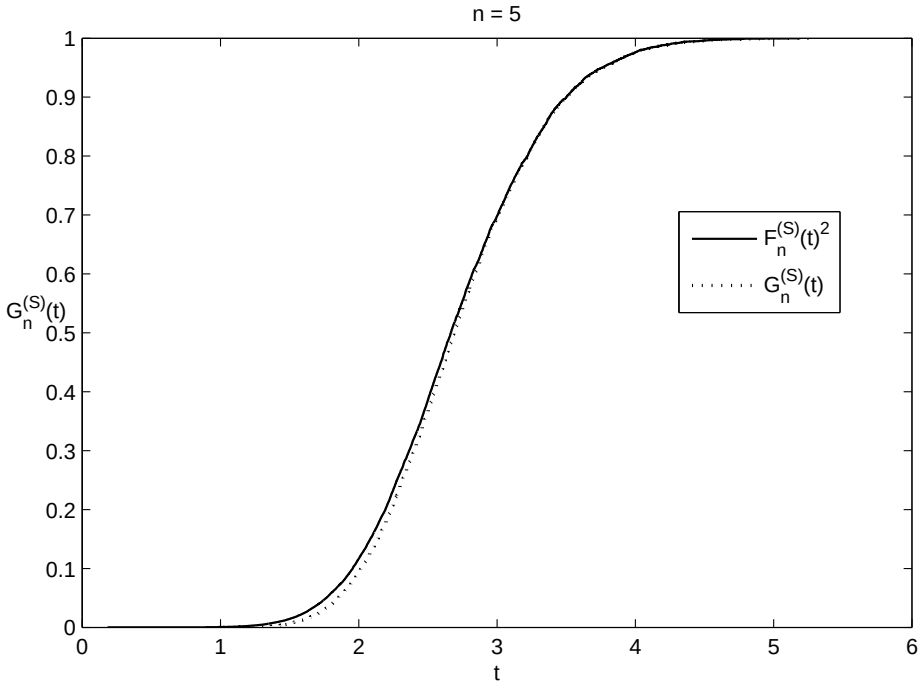

Figure 2-6: *Comparison between* $F_n^{(S)}(t)^2$ *and* $G_n^{(S)}(t)$ *for* $n = 5$.

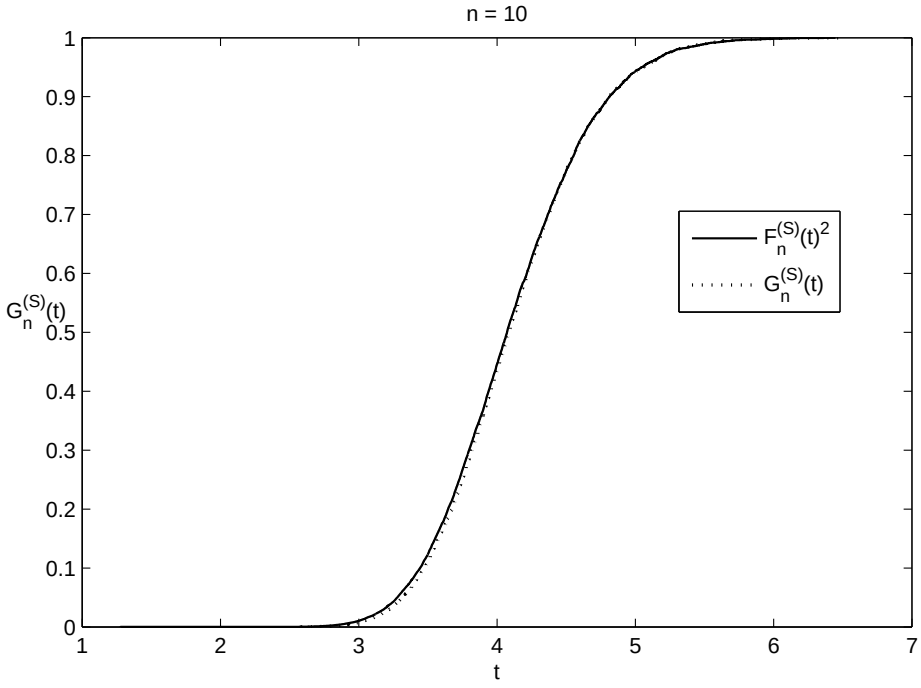

Figure 2-7: *Comparison between* $F_n^{(S)}(t)^2$ *and* $G_n^{(S)}(t)$ *for* $n = 10$.

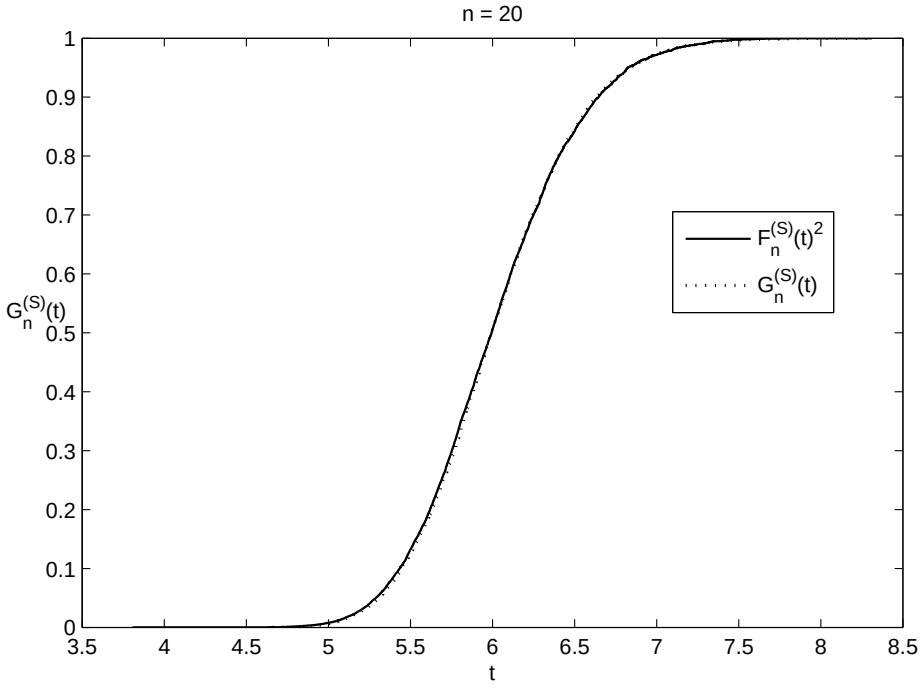

Figure 2-8: *Comparison between* $F_n^{(S)}(t)^2$ *and* $G_n^{(S)}(t)$ *for* $n = 20$.

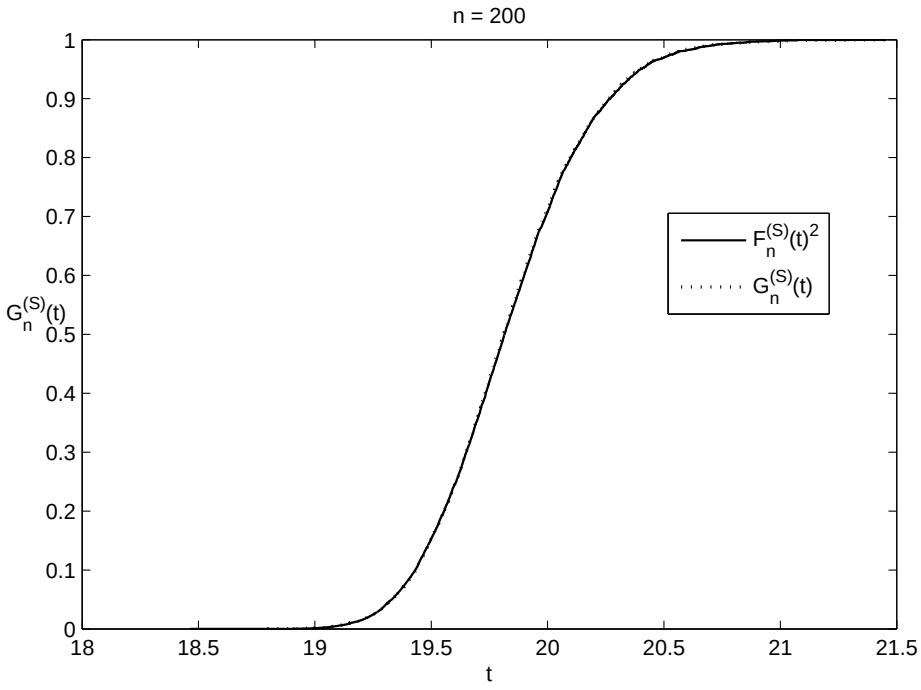

Figure 2-9: *Comparison between* $F_n^{(S)}(t)^2$ *and* $G_n^{(S)}(t)$ *for* $n = 200$.

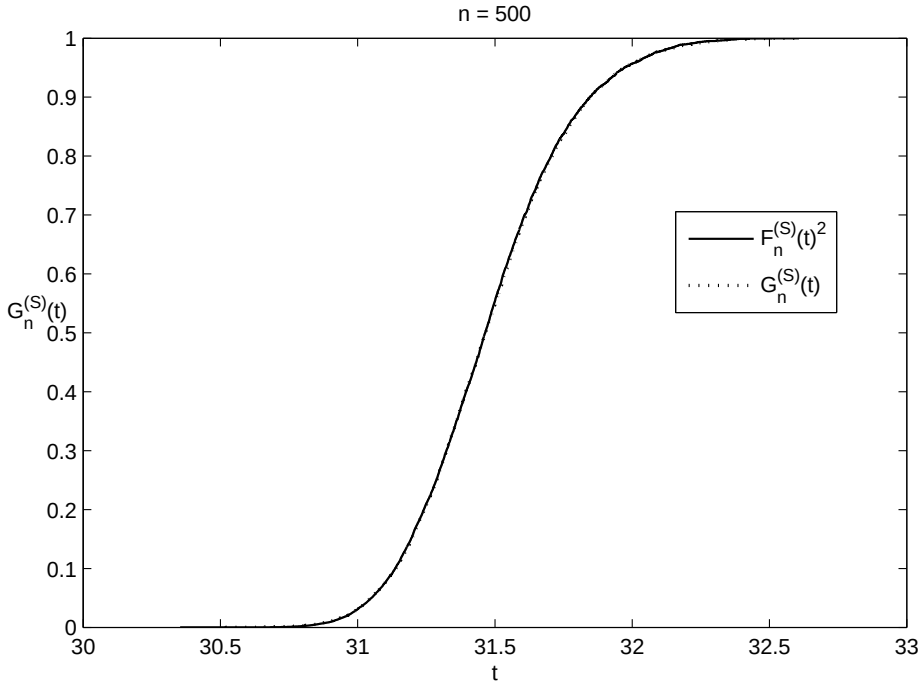

Figure 2-10: *Comparison between* $F_n^{(S)}(t)^2$ *and* $G_n^{(S)}(t)$ *for* $n = 500$.

An implication of Conjecture 2.3.1, and equation (2.20) in particular, is that we may make a minor alteration to Program 2.2.1, to produce an approximation to $G_n(t)$, which we denote as $G_n^{(P)}(t)$. (The superscript (P) reflects the fact that $G_n^{(P)}(t)$, similarly to $F_n^{(P)}(t)$, is obtained from the numerical solution of the system of Painlevé II ODEs, (2.11), with initial conditions (2.12).) This can be done in the following way: Once we have obtained $F(s)$, the numerical solution to (2.11) with initial conditions (2.12), we let $G(s) := F(s)^2$. Then, using the relation between s and t, given in (2.5), we get $G_n^{(P)}(t)$ from $G(s)$ by simply letting

$$G_n^{(P)}(t) := G(s).$$

In fact, as we already saw in the experimental justification of Conjecture 2.3.1, the convergence in (2.20) is so fast that the sign "$\overset{n}{=}$" there may be treated as an equality. Therefore, one should expect that $G_n^{(P)}(t)$ approximates $G_n(t)$ as well as $F_n^{(P)}(t)$ approximates $F_n(t)$. The latter is indicated in Figures 2-11, 2-12, 2-13, 2-14 and 2-15 below.

So far we conjectured that the events $\{\beta_1 < t\}$ and $\{\beta_n < t\}$ become "less dependent" as n increases (see Conjecture 2.3.1 above). As a consequence, we derived an expression for the c.d.f. of $\|B\|_2$, $G_n(t)$, via the c.d.f. of β_n, $F_n(t)$. The relation

between the two distribution functions was given in (2.20). We tested numerically Conjecture 2.3.1 by testing its equivalent statement, (2.20), and the results showed that it is reasonable to assume that in fact $G_n(t) = F_n(t)^2$ for $n \geq 20$ (see Figures 2-6, 2-7, 2-8, 2-9 and 2-10). We further noted that an important consequence of (2.20) is that we could use Program 2.2.1 to calculate an approximation of $G_n(t)$, which we denoted as $G_n^{(P)}(t)$. So we are now in a position to give the slight modification of Program 2.2.1, which is needed for the calculation of $G_n^{(P)}(t)$, and test how well $G_n^{(P)}(t)$ approximates $G_n(t)$. Before we do that, let us first introduce the probability density function (p.d.f.) of $\|B\|_2$. The function $G_n(t)$ was defined as $\|B\|_2$'s cumulative distribution function (c.d.f.). Therefore the p.d.f. of $\|B\|_2$ will be

$$g_n(t) := \frac{d}{dt} G_n(t).$$

Similarly to $G_n^{(P)}(t)$, we denote by $g_n^{(P)}(t)$ the p.d.f. of $\|B\|_2$, obtained from the numerical solution to the system of ODEs (2.11) with initial conditions (2.12) (see Remark A.1.3 for more details about the relation between $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$). Thus, with our next experiment we shall also test how well $g_n^{(P)}(t)$ approximates $g_n(t)$.

Program 2.3.1 below is a slightly modified version of Program 2.2.1 from (Edelman and Persson, 2005). Our modification produces $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ as outputs. In Program 2.3.1 $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ are denoted as `G` and `g`, respectively, where the variable t is obtained from s by the transformation

```
t = sqrt(2*n) + s/(n^{1/6}*sqrt(2))
```

as given in (2.5).

Program 2.3.1.

```
n = 500;

deq = inline('[y(2); s*y(1) + 2*y(1)^3; y(4); y(1)^2; -y(1)]', ...
    's', 'y');

s0 = 5;
sn = -8;
sspan = linspace(s0, sn, 1000);

y0 = [airy(s0); airy(1,s0);...
```

```
    quadl(inline('(x - s0).*airy(x).^2', 'x', 's0'), s0, 20, 1e-25, 0, s0); ...
    airy(s0)^2; quadl(inline('airy(x)'), s0, 20, 1e-18)];

opts = odeset('reltol', 1e-13, 'abstol', 1e-15);
[s, y] = ode45(deq, sspan, y0, opts);

t = sqrt(2*n) + s/(n^(1/6)*sqrt(2));

G = exp(-y(:, 5) - y(:, 3));

g = n^(1/6)*sqrt(2)*(y(:, 1) - y(:, 4)).*G;
```

Remark 2.3.1. *The last line of Program 2.3.1 reflects the fact that the solution to the system of ODEs* (2.11) *with initial conditions* (2.12)*, $G(s)$, is obtained in terms of the variable s. Then, in order to obtain $G_n^{(P)}(t)$, we let $G_n^{(P)}(t) := G(s(t))$, where*

$$s(t) = n^{1/6}\sqrt{2}(t - \sqrt{2n}),$$

which is the inverse of the transformation (2.5)*. Therefore*

$$g_n^{(P)}(t) = \frac{d}{dt}G_n^{(P)}(t) = \frac{d}{dt}G(s(t)) = G'(s(t))s'(t) = n^{1/6}\sqrt{2}G'(s),$$

where it is easy to show from (2.13) *that*

$$G'(s) = 2F(s)F'(s) = -(I'(s) + J'(s))F(s)^2 = (q(s) - I'(s))G(s).$$

The next step is to compare $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$, obtained by solving the system of ODEs (2.11) with initial conditions (2.12), with the functions $G_n^{(S)}(t)$ and $g_n^{(S)}(t)$, which are derived by simulation, using Program 2.3.2 below. The program creates 10 000 samples of the matrix B and for each sample it stores the value of $\|B\|_2$ as an entry of the vector `NormB`. Then the built-in MATLAB functions `ecdf` and `ksdensity` are applied to the entries of the vector `NormB` to produce the c.d.f. $G_n^{(S)}(t)$ and the p.d.f. $g_n^{(S)}(t)$, which in Program 2.3.2 are denoted as `Gsim` and `gsim`, respectively.

Program 2.3.2.

```
n = 500;
nrep = 1e4;
NormB = zeros(1, nrep);
```

```
matlabpool open 4
parfor(ii = 1:nrep)
    B = randn(n);
    B = (B + B')/2;
    NormB(ii) = norm(B);
end
matlabpool close

[Gsim, tsim] = ecdf(NormB);
gsim = ksdensity(NormB, tsim);
```

In Figures 2-11, 2-12, 2-13, 2-14 and 2-15, on the left side of each figure, we have plotted $g_n^{(P)}(t)$ and $g_n^{(S)}(t)$ and on the right side there are $G_n^{(P)}(t)$ and $G_n^{(S)}(t)$. Beneath each of the figures we have also given the value of $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)|$ for comparison with $\max_t |F_n^{(P)}(t) - F_n^{(S)}(t)|$ from Figures 2-1, 2-2, 2-3, 2-4 and 2-5.

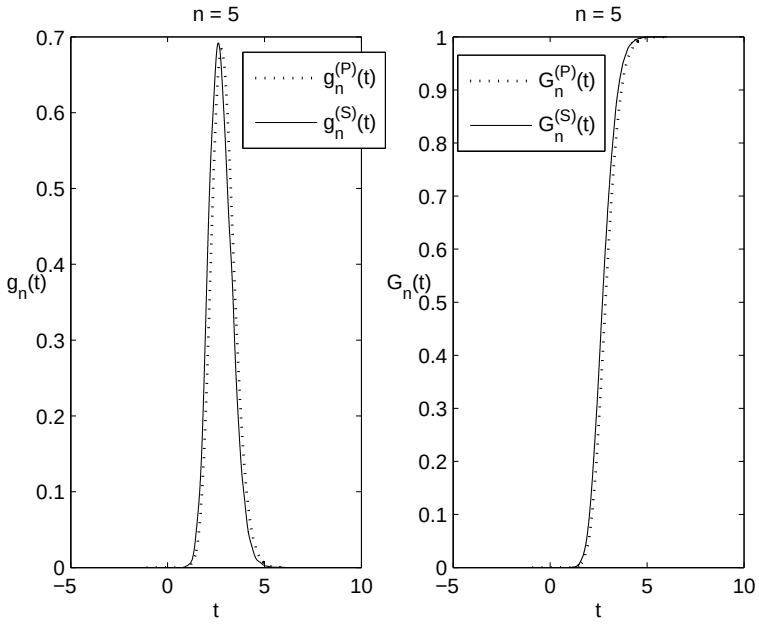

Figure 2-11: *On the left figure we compare* $g_n^{(P)}(t)$ *with* $g_n^{(S)}(t)$ *and on the right one we compare* $G_n^{(P)}(t)$ *with* $G_n^{(S)}(t)$ *for* $n = 5$. *Here* $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)| = 0.1087$.

If we consider $G_n^{(S)}(t)$ and $g_n^{(S)}(t)$ as "reliable representatives" of $G_n(t)$ and $g_n(t)$

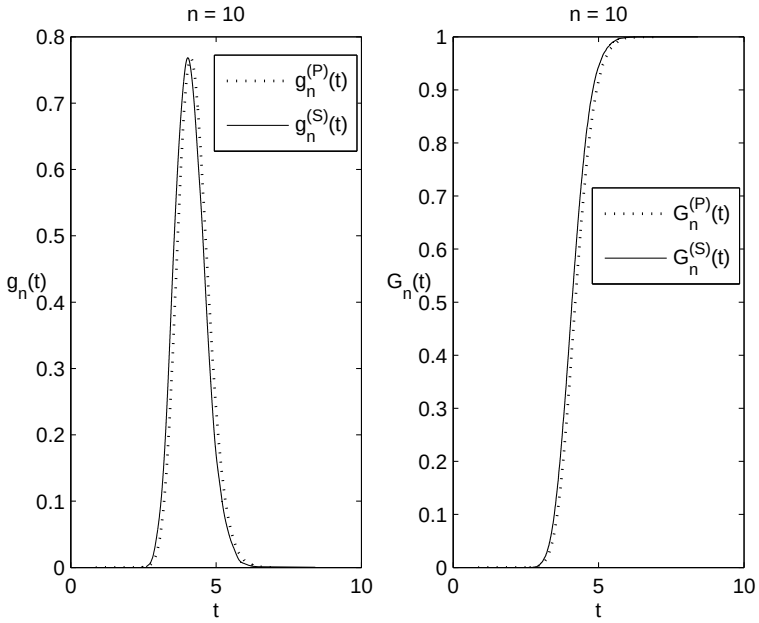

Figure 2-12: *On the left figure we compare* $g_n^{(P)}(t)$ *with* $g_n^{(S)}(t)$ *and on the right one we compare* $G_n^{(P)}(t)$ *with* $G_n^{(S)}(t)$ *for* $n = 10$. *Here* $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)| = 0.0889$.

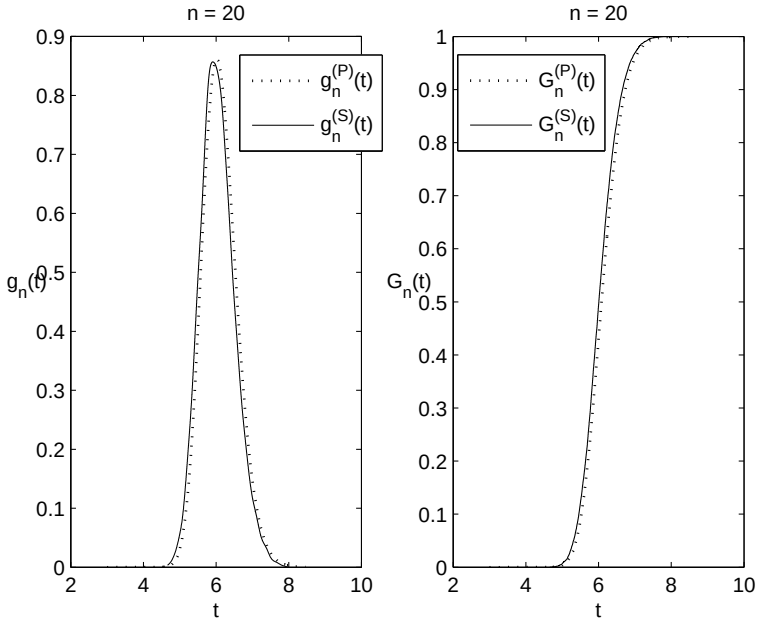

Figure 2-13: *On the left figure we compare* $g_n^{(P)}(t)$ *with* $g_n^{(S)}(t)$ *and on the right one we compare* $G_n^{(P)}(t)$ *with* $G_n^{(S)}(t)$ *for* $n = 20$. *Here* $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)| = 0.0648$.

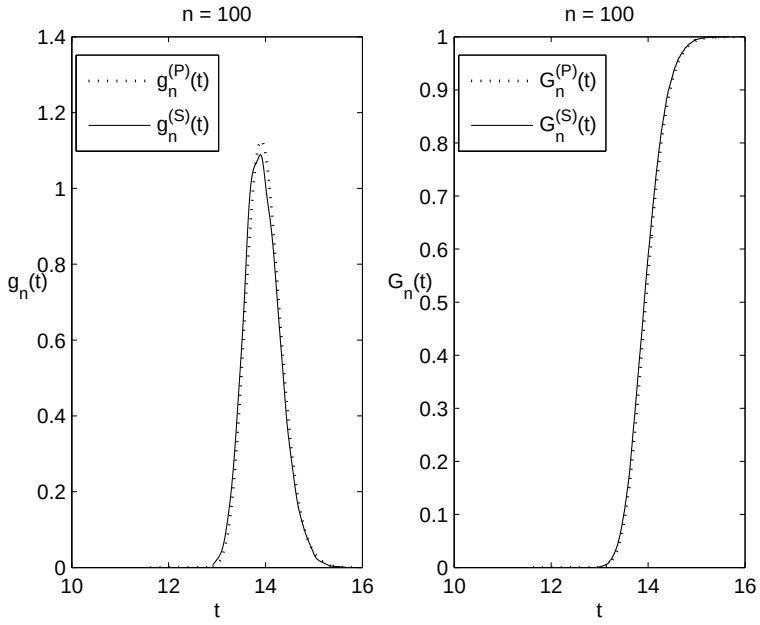

Figure 2-14: *On the left figure we compare* $g_n^{(P)}(t)$ *with* $g_n^{(S)}(t)$ *and on the right one we compare* $G_n^{(P)}(t)$ *with* $G_n^{(S)}(t)$ *for* $n = 100$. *Here* $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)| = 0.0437$.

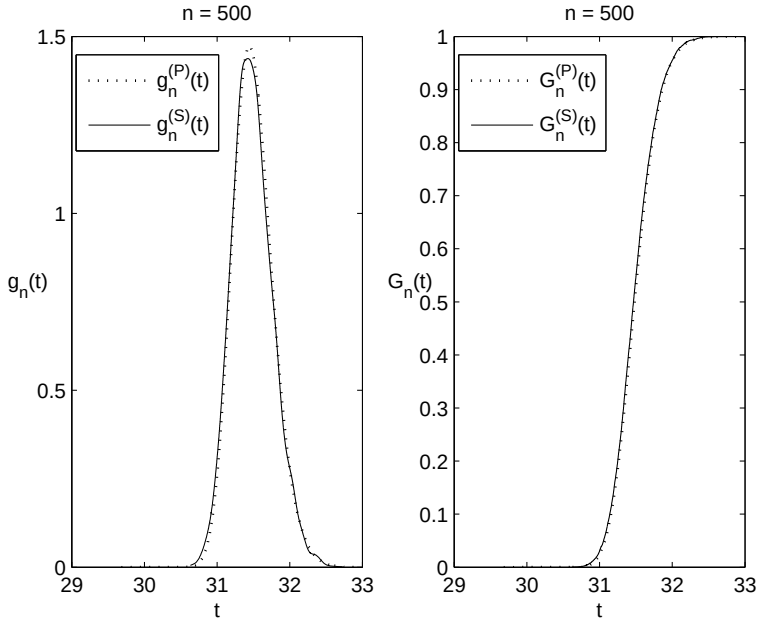

Figure 2-15: *On the left figure we compare* $g_n^{(P)}(t)$ *with* $g_n^{(S)}(t)$ *and on the right one we compare* $G_n^{(P)}(t)$ *with* $G_n^{(S)}(t)$ *for* $n = 500$. *Here* $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)| = 0.024$.

respectively, that is, if we assume

$$G_n(t) \approx G_n^{(S)}(t) \qquad \text{and} \qquad g_n(t) \approx g_n^{(S)}(t),$$

we can infer that $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ are good approximations to $G_n(t)$ and $g_n(t)$ respectively, since the errors, $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)|$, decrease with n. We can also see that, for the cases of n considered here, the values of $\max_t |G_n^{(P)}(t) - G_n^{(S)}(t)|$ look "very similar" to those of $\max_t |F_n^{(P)}(t) - F_n^{(S)}(t)|$ (see captions of Figures 2-1, 2-2, 2-3, 2-4 and 2-5 for comparison). In other words, $G_n^{(P)}(t)$ approximates $G_n(t)$ as well as $F_n^{(P)}(t)$ approximates $F_n(t)$, as suggested above. This can be taken as a clear indicator that the assumption in (2.17) makes sense numerically and Conjecture 2.3.1, and thus (2.20), are correct. Hence, as a conclusion, $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ are good approximations to $G_n(t)$ and $g_n(t)$, respectively. Also, when n is large, the values of $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ can be obtained much faster, by solving numerically (2.11) with initial conditions (2.12), than the values of $G_n^{(S)}(t)$ and $g_n^{(S)}(t)$, by simulating $\|B\|_2$. In fact, for very large values of n, due to restriction in computer memory, it is impossible to find $\|B\|_2$, and thus $G_n^{(S)}(t)$ and $g_n^{(S)}(t)$, while the complexity of calculating $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ doesn't change with n. Moreover, according to the limit in (2.3) and assuming Conjecture 2.3.1, as n increases, $G_n(t)$ and $g_n(t)$ converge to $G(s)$ and $g(s) := G'(s)$ respectively, and therefore $G_n^{(P)}(t)$ and $g_n^{(P)}(t)$ become better approximations of $G_n(t)$ and $g_n(t)$.

2.4. Extensions to Conjecture 2.3.1 for other classes of matrices

Here we list extensions to Conjecture 2.3.1 for other classes of matrices. They are given as Conjectures 2.4.1 and 2.4.2 in the text below. We have tested these conjectures numerically on GOE and 4 additional classes of matrices, including random Laplacian matrices (see Definition 2.4.1 and the definition of random Laplacian matrices afterwards). We also present a proof of a weaker version of Conjecture 2.4.1 (see Theorem 2.4.1).

GOE matrices (Here we simply recall Definition 2.2.1.) This is the class of matrices, which we considered so far in this chapter. We recall that $B \in \mathbb{R}^{n\times n}$ is a GOE matrix, if it is symmetric and its entries above and on the main diagonal, B_{ij}, $1 \le i \le j \le n$, are independent random variables, such that

$$B_{ii} \in \mathcal{N}(0,1), \quad 1 \le i \le n, \qquad \text{and} \qquad B_{ij} \in \mathcal{N}\left(0, \frac{1}{2}\right), \quad 1 \le i < j \le n.$$

Also, since B is symmetric, $B_{ij} = B_{ji}$.

Scaled GOE matrices The matrix $B \in \mathbb{R}^{n\times n}$ is said to be a Scaled GOE matrix, if it is symmetric and its entries above and on the main diagonal, B_{ij}, $1 \leq i \leq j \leq n$, are independent random variables, such that

$$B_{ii} \in \mathcal{N}\left(0, \frac{2}{n}\right), \quad 1 \leq i \leq n, \qquad \text{and} \qquad B_{ij} \in \mathcal{N}\left(0, \frac{1}{n}\right), \quad 1 \leq i < j \leq n.$$

In other words, if B is a GOE matrix, then the matrix $\sqrt{\frac{2}{n}}B$ is a Scaled GOE matrix.

Uniform matrices These are random symmetric matrices, whose entries above and on the main diagonal are independent, identically distributed random variables, uniformly distributed over the interval $(-0.5, 0.5)$.[2]

Bernoulli matrices This is the class of random symmetric matrices, whose entries above and on the main diagonal are independent, identically distributed random variables, taking the values -0.5 or 0.5 with probability 0.5.[3]

Definition 2.4.1 (**Generalised Laplacian matrices**)**.** Let $L \in \mathbb{R}^{n\times n}$ be symmetric matrix, whose diagonal entries, L_{ii}, satisfy

$$L_{ii} = -\sum_{j=1, j\neq i}^{n} L_{ij} \tag{2.21}$$

for $1 \leq i \leq n$. Then L is called a generalised Laplacian matrix.

Remark 2.4.1. *Note the difference between Definition 2.4.1 and Definition 1.2.12 of Laplacian matrix, where we required the off-diagonal entries to be non-positive. We remove that requirement in Definition 2.4.1, because we shall use generalised Laplacian matrices to perturb Laplacian matrices of graphs, in order to model positive, as well as negative, perturbations to the weights of the edges in the graph.*

Random Laplacian matrices In general, random Laplacian matrices are symmetric Laplacian matrices, whose entries above the main diagonal are i.i.d. random variables and the relation between diagonal and off-diagonal elements is given by (2.21). However, in this chapter we shall only consider random Laplacian matrices, whose off diagonal elements are distributed $\mathcal{N}\left(0, \frac{1}{n}\right)$, where n is the size of the matrix.

[2] See Definition A.1.15, where we define *Uniform distribution.*
[3] See Definition A.1.14, where we define *Bernoulli distribution.*

We are now ready to state the conjectures.

Conjecture 2.4.1. *Let $B \in \mathbb{R}^{n\times n}$ be a random symmetric matrix, whose entries above the main diagonal, B_{ij}, $1 \leq i < j \leq n$, are independent, identically distributed random variables, such that B_{ij} and $-B_{ij}$ have the same distribution. Also, let the diagonal elements of B, B_{ii}, $1 \leq i \leq n$, be independent identically distributed (i.i.d.) random variables, independent from B_{ij}, $1 \leq i < j \leq n$, and such that the distribution of B_{ii} is the same as that of $-B_{ii}$. Then, if $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ are the eigenvalues of B, we have*

$$\mathbb{P}[\|B\|_2 < t] \overset{n}{=} \mathbb{P}[\beta_n < t]^2. \tag{2.22}$$

Remark 2.4.2. *In the statement of Conjecture 2.4.1 the diagonal elements of B may or may not have the same distribution as the elements of B above the diagonal. The class of matrices, for which Conjecture 2.4.1 holds, includes GOE matrices, Scaled GOE matrices, Uniform matrices and Bernoulli matrices from the classes of matrices listed above. However, it doesn't include random Laplacian matrices, since the diagonal entries of the latter depend on their off-diagonal elements (see* (2.21)*). This is why we have stated a separate conjecture for random Laplacian matrices (see Conjecture 2.4.2).*

Although we were not able to prove Conjecture 2.4.1 rigorously in such a general form, we managed to prove a weaker result, given in the following corollary. The proof came as a result of discussions with Rob Scheichl and Alex Cox, both from the Department of Mathematical Sciences, University of Bath.

Theorem 2.4.1. *Let the matrix B satisfy the statement of Conjecture 2.4.1 and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ be its eigenvalues. Suppose $\beta_n \xrightarrow{\mathcal{D}} c$, where $c > 0$ is some constant. Then*

$$\mathbb{P}[\|B\|_2 < t] \overset{n}{=} \mathbb{P}[\beta_n < t]^2 \tag{2.23}$$

for all $t \in \mathbb{R}$ and thus, $\|B\|_2 \xrightarrow{\mathcal{D}} c$.

Remark 2.4.3. *Before we present the proof of this corollary, we recall that $\xrightarrow{\mathcal{D}}$ denotes* convergence in distribution *(for more details see Definition A.1.16)*

Proof. Firstly, we shall show that $|\beta_n| \xrightarrow{\mathcal{D}} c$, as $n \to \infty$. Let us take some $t > c$. Then

$$\mathbb{P}[|\beta_n| < t] = \mathbb{P}[-t < \beta_n < t] = \mathbb{P}[\beta_n < t] - \mathbb{P}[\beta_n \leq -t] \to 1, \quad \text{as } n \to \infty.$$

If, on the other hand, we take some $0 < t < c$, we obtain

$$\mathbb{P}[|\beta_n| < t] = \mathbb{P}[-t < \beta_n < t] \leq \mathbb{P}[\beta_n < t] \to 0, \quad \text{as } n \to \infty.$$

Hence $|\beta_n| \xrightarrow{\mathcal{D}} c$.

Secondly, we shall show that $|\beta_1| \xrightarrow{\mathcal{D}} c$, as $n \to \infty$. Since the elements of the matrices B and $-B$ have exactly the same distributions, we can conclude that β_n has the same distribution as $-\beta_1$ (since the latter is the largest eigenvalue of $-B$). Therefore $-\beta_1 \xrightarrow{\mathcal{D}} c$, as $n \to \infty$ and hence, as we proved above $|\beta_1| \xrightarrow{\mathcal{D}} c$, as $n \to \infty$.

Lastly, we shall show that $\|B\|_2 \xrightarrow{\mathcal{D}} c$. Since, by definition, $\|B\|_2 = \max\{|\beta_1|, |\beta_n|\}$, we have

$$\mathbb{P}[\|B\|_2 < t] = \mathbb{P}[|\beta_1| < t, |\beta_n| < t]$$

for all $t \in \mathbb{R}$. Let us take $0 < t < c$. Then we obtain

$$\mathbb{P}[\|B\|_2 < t] \leq \mathbb{P}[|\beta_n| < t] \to 0, \quad \text{as } n \to \infty.$$

Now let us take $t > c$. We have

$$\begin{aligned} \mathbb{P}[\|B\|_2 < t] &= \mathbb{P}[|\beta_1| < t, |\beta_n| < t] \\ &= 1 - \mathbb{P}[|\beta_1| \geq t, |\beta_n| < t] - \mathbb{P}[|\beta_1| < t, |\beta_n| \geq t] - \mathbb{P}[|\beta_1| \geq t, |\beta_n| \geq t] \\ &\geq 1 - \mathbb{P}[|\beta_1| \geq t] - \mathbb{P}[|\beta_n| \geq t] - \mathbb{P}[|\beta_1| \geq t] \to 1, \quad \text{as } n \to \infty. \end{aligned}$$

Therefore, we can conclude that

$$\mathbb{P}[\|B\|_2 < t] = \begin{cases} 1 & \text{if } t > c \\ 0 & \text{if } t < c, \end{cases}$$

or in other words, $\|B\|_2 \xrightarrow{\mathcal{D}} c$, as $n \to \infty$, which also implies that (2.23) holds. □

Remark 2.4.4. *Essentially, the proof of Theorem 2.4.1 proves that if two sequences of random variables, $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$, are such that the distributions of X_n and $-Y_n$ are the same for each $n \in \mathbb{N}$, and $X_n \xrightarrow{\mathcal{D}} c$ for some $c > 0$, then the random variables $\max\{|X_n|, |Y_n|\} \xrightarrow{\mathcal{D}} c$, as $n \to \infty$. Therefore, in the statement of Theorem 2.4.1 we only need the requirement that the distributions of the elements of B and $-B$ are the same and $\beta_n \xrightarrow{\mathcal{D}} c$ for some $c > 0$.*

Remark 2.4.5. *As we shall see later, Theorem 2.4.1 implies that if B is a Scaled GOE matrix, then $\|B\|_2 \xrightarrow{\mathcal{D}} 2$, as $n \to \infty$. This result is stated as Corollary 2.4.1.*

Remark 2.4.6. *From the proof of Theorem 2.4.1 we can see that in fact one can prove that*

$$\mathbb{P}[\|B\|_2 < t] \overset{n}{=} \mathbb{P}[\beta_n < t]^a \tag{2.24}$$

for any constant $a > 0$. Interesting questions (which we are unable to answer at this moment) arising from (2.24) *are: Is there a value of a, such that the convergence in* (2.24) *is fastest? If yes, what is that value? In* (2.23) *we took $a = 2$, because this seemed a "natural" to us. We explain this choice in the following way: When the size of the matrix, n, increases, the number of eigenvalues between β_1 and β_n increases and therefore it is "natural" to assume that they become less dependent. Thus, their relation, in the limit as $n \to \infty$, should resemble that of independent random variables. However, the latter is only an argument based on our intuition, which we haven't been able to prove yet.*

Conjecture 2.4.2. *Let $L \in \mathbb{R}^{n\times n}$ be a random Laplacian matrix. Then, if $l_1 \leq l_2 \leq \cdots \leq l_n$ are the eigenvalues of L, we have*

$$\mathbb{P}[\|L\|_2 < t] \stackrel{n}{=} \mathbb{P}[l_n < t]^2. \tag{2.25}$$

Justification of Conjectures 2.4.1 and 2.4.2 (by experiments). From the proof of Theorem 2.4.1 we know that in the case of Scaled GOE matrix the convergence (2.23) holds. However, here we check something similar to uniform convergence of disitribution functions (when the domains are discrete sets of points), which is stronger than the pointwise convergence in (2.22), (2.23) and (2.25). We perform the same experiment, based on simulations, as for Conjecture 2.3.1, with the only difference that now we consider 5 different classes of matrices (described at the beginning of this section): GOE matrices, Scaled GOE matrices, Uniform matrices, Bernoulli matrices and Laplacian matrices. The values of n which we consider are $n = 10, 20, 100, 200$ and 500. For each n among these values and for each of the 5 classes of random matrices listed above we find $F_n^{(S)}(t)$ and $G_n^{(S)}(t)$ by simulation (see the justification of Conjecture 2.3.1 for more details about the simulations). We then compare $F_n^{(S)}(t)^2$ and $G_n^{(S)}(t)$ graphically (see Figures 2-16, 2-17, 2-18, 2-19, 2-20, 2-21, 2-22 and 2-23) and also calculate the value of $\max_t |G_n^{(S)}(t) - F_n^{(S)}(t)^2|$ (see Table 2.2) to check (and compare) the convergence in (2.22) and in (2.25) for the different classes of random matrices.

From Figures 2-16, 2-17, 2-18, 2-19, 2-20, 2-21, 2-22 and 2-23 and from the results in Table 2.2 we can see that the numerics agree well with the statements of Theorem 2.4.1 and Conjectures 2.4.1 and 2.4.2. Although different classes of matrices give different values for $\max_t |G_n^{(S)}(t) - F_n^{(S)}(t)^2|$ when n is small, we can see (from Table 2.2) that the values of $\max_t |G_{500}^{(S)}(t) - F_{500}^{(S)}(t)^2|$ are already of similar magnitude.

From the numerical results given in Table 2.2, the class of matrices which seems to converge faster than the others is that of Scaled GOE matrices (given as Sc. GOE in Table 2.2). However, the classes of Bernoulli and Laplacian random matrices start

with $\max_t |G_{10}^{(S)}(t) - F_{10}^{(S)}(t)^2| \approx 0.07$ and both reach $\max_t |G_{500}^{(S)}(t) - F_{500}^{(S)}(t)^2| \approx 0.0065$, which can be interpreted as a good sign that the conjectures hold for these two classes of matrices.

As a conclusion, taking into account the results in Figures 2-16, 2-17, 2-18, 2-19, 2-20, 2-21, 2-22 and 2-23 and Table 2.2, we can say that there is a strong evidence that the claims of Conjectures 2.4.1 and 2.4.2 are true, probably with the only exception of $\max_t |G_{500}^{(S)}(t) - F_{500}^{(S)}(t)^2| = 0.0093$ for GOE matrices (see Remark 2.4.7 for further discussion).

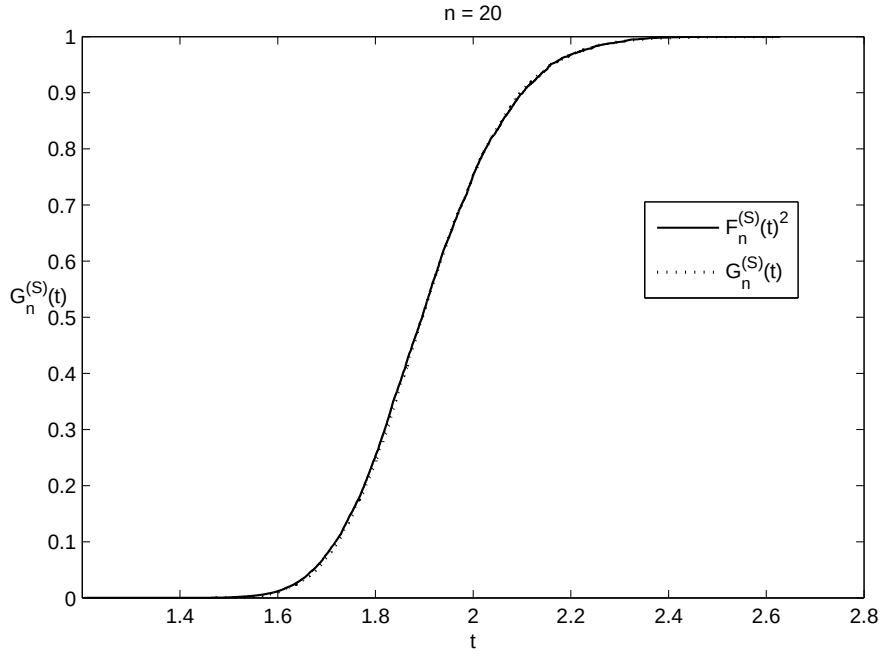

Figure 2-16: *Comparison between $G_n^{(S)}(t)$ and $F_n^{(S)}(t)^2$ for $n = 20$ for Scaled GOE random matrices.*

□

Remark 2.4.7. *This remark comments upon the "unexpected" result for* $\max_t |G_{500}^{(S)}(t) - F_{500}^{(S)}(t)^2|$ *in Table 2.2. Further experiments with the GOE class of matrices showed that the value of* $\max_t |G_n^{(S)}(t) - F_n^{(S)}(t)^2|$ *is very sensitive to the number of simulations we perform and improves significantly when we increase that number. For example, when we increased the number of simulations from* 10 000*, which we have used for all experiments here, to* 100 000*, and to* 1 000 000*, the value of* $\max_t |G_{500}^{(S)}(t) - F_{500}^{(S)}(t)^2|$ *became* 0.0022 *and* 0.0013*, respectively. Another difference which occurred was that when we performed* 10 000 *simulations, we obtained the numerical values of* β_1 *between* -32.8395 *and* -30.2529*, while those of* β_n *were between* 30.4053 *and* 32.7874*. In other words,*

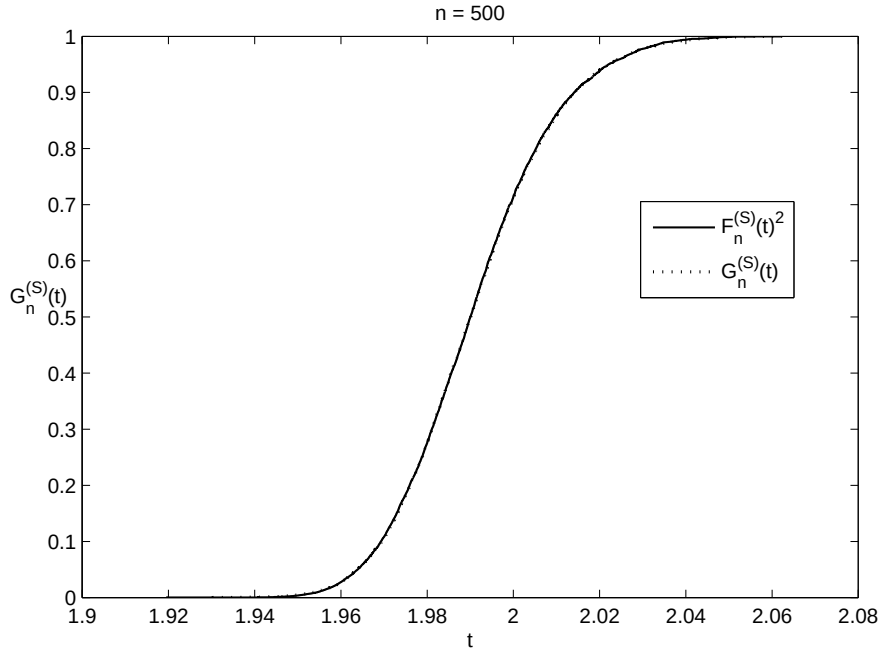

Figure 2-17: *Comparison between $G_n^{(S)}(t)$ and $F_n^{(S)}(t)^2$ for $n = 500$ for Scaled GOE random matrices.*

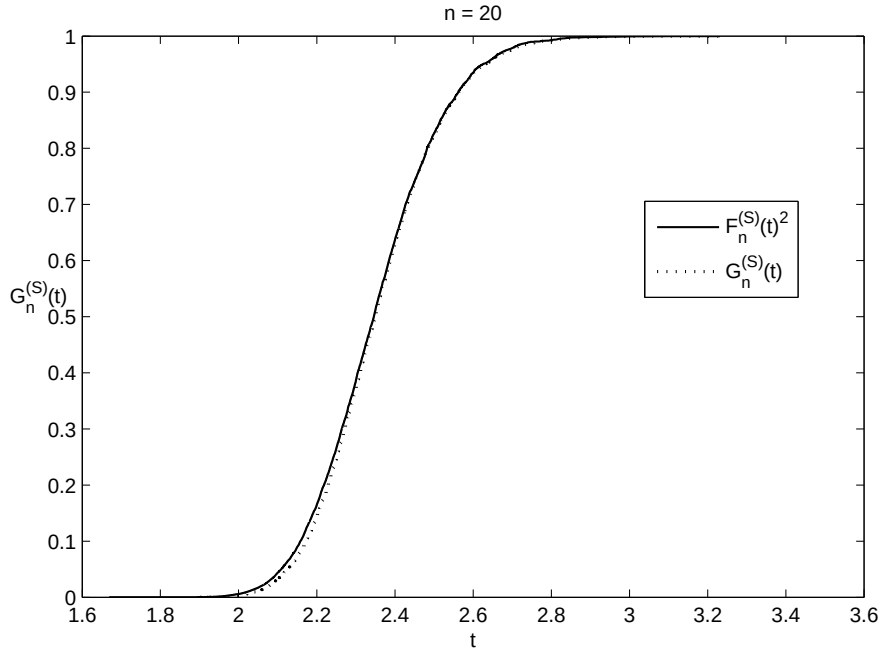

Figure 2-18: *Comparison between $G_n^{(S)}(t)$ and $F_n^{(S)}(t)^2$ for $n = 20$ for Uniform random matrices.*

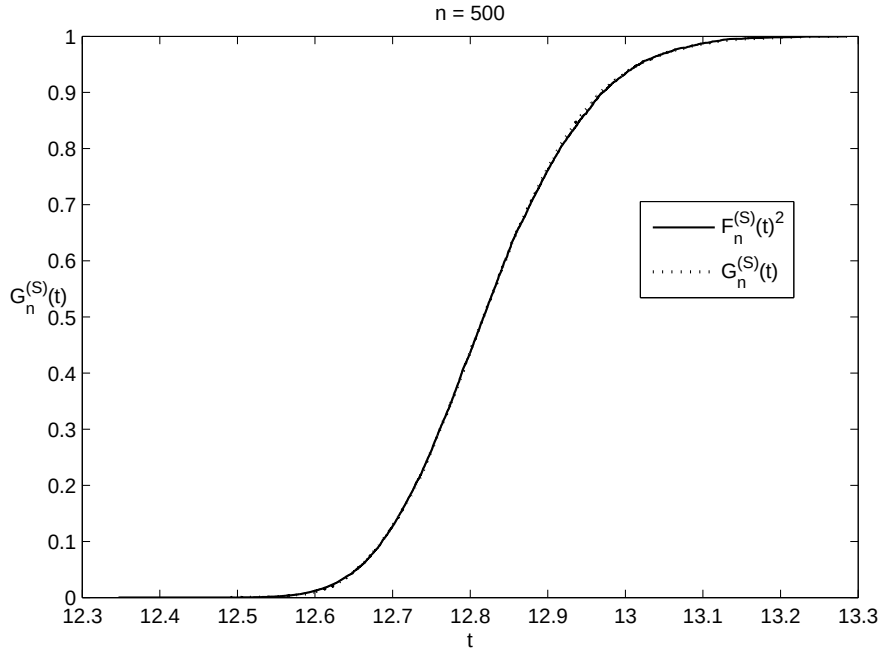

Figure 2-19: *Comparison between* $G_n^{(S)}(t)$ *and* $F_n^{(S)}(t)^2$ *for* $n = 500$ *for Uniform random matrices.*

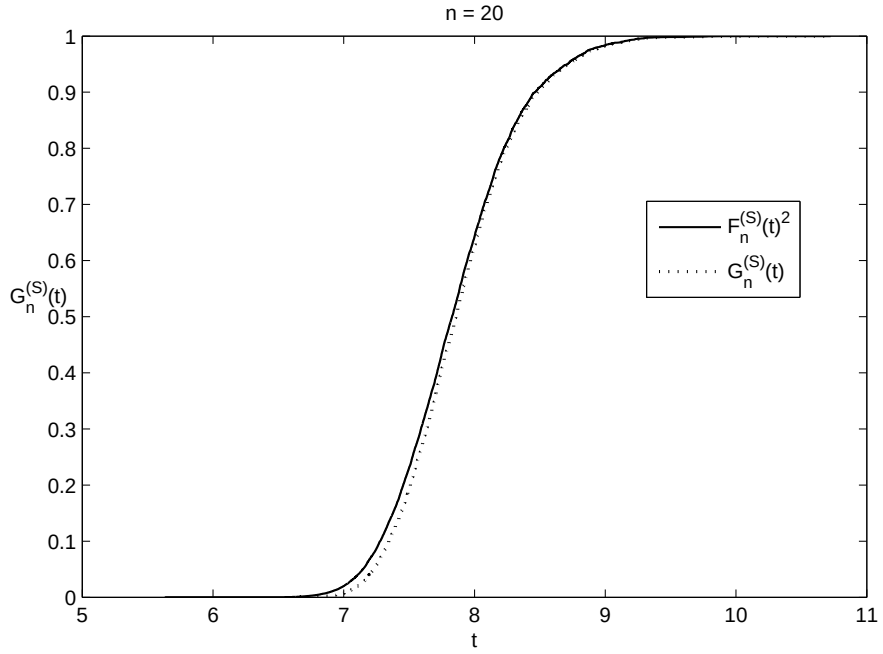

Figure 2-20: *Comparison between* $G_n^{(S)}(t)$ *and* $F_n^{(S)}(t)^2$ *for* $n = 20$ *for Bernoulli random matrices.*

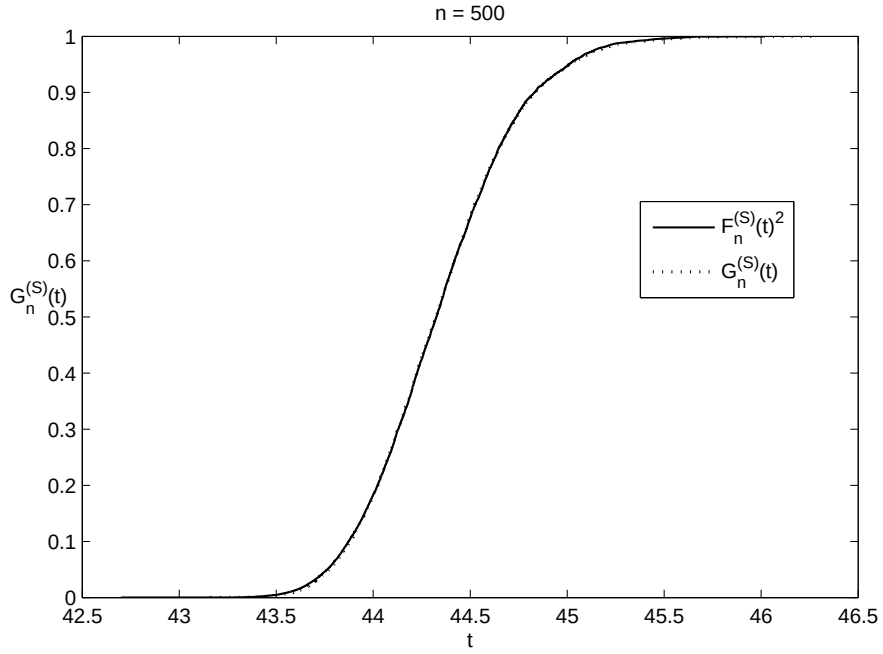

Figure 2-21: *Comparison between* $G_n^{(S)}(t)$ *and* $F_n^{(S)}(t)^2$ *for* $n = 500$ *for Bernoulli random matrices.*

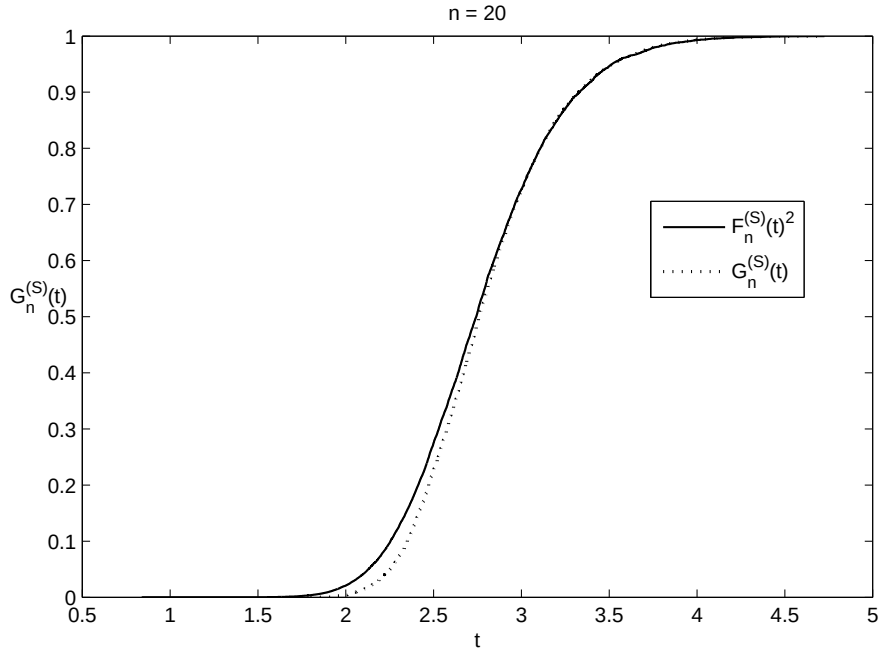

Figure 2-22: *Comparison between* $G_n^{(S)}(t)$ *and* $F_n^{(S)}(t)^2$ *for* $n = 20$ *for Laplacian random matrices.*

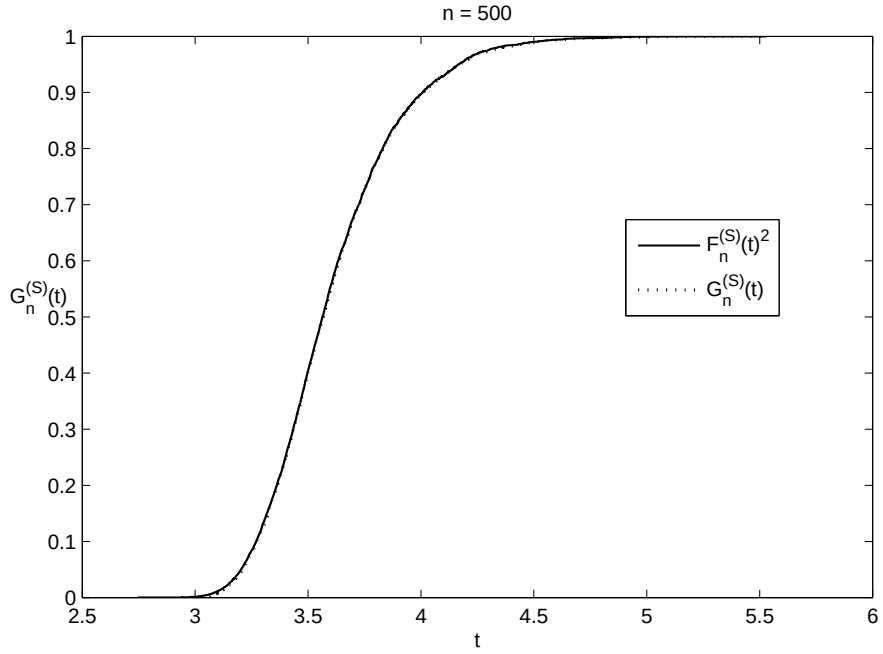

Figure 2-23: *Comparison between* $G_n^{(S)}(t)$ *and* $F_n^{(S)}(t)^2$ *for* $n = 500$ *for Laplacian random matrices.*

$|\beta_1| \in [30.2529, 32.8395]$ *and* $|\beta_n| \in [30.4053, 32.7874]$. *With* 100 000 *simulations these intervals became* $[30.0361, 32.9431]$ *and* $[30.0368, 33.038]$ *and with* 1 000 000 *simulations,* $[29.9294, 33.2384]$ *and* $[29.9744, 33.1182]$, *for* $|\beta_1|$ *and* $|\beta_n|$, *respectively. In theory, the intervals in which* $|\beta_1|$ *and* $|\beta_n|$ *lie should be identical, as both random variables have the same distribution. However, in a numerical experiment the values of the empirical c.d.f.'s of* $|\beta_1|$ *and* $|\beta_n|$ *are obtained only for a discrete set of points from the empirical intervals given above. For any point outside those intervals we let the c.d.f.'s equal zero, if the point is less than the left end point of the interval and one, if it is greater than the right end point. Since* $\|B\|_2 = \max\{|\beta_1|, |\beta_n|\}$, *if in our set of simulations it has turned out that, say* $|\beta_1| > |\beta_n|$ *in more cases than* $|\beta_1| \leq |\beta_n|$, *then this would "distort" the distribution of* $\|B\|_2$. *Similarly, if, for example, the empirical interval for* $|\beta_1|$ *lies "more to the left", compared to that of* $|\beta_n|$, *then the distribution of* $\|B\|_2$ *is also "distorted". The extent to which that "distortion" occurs depends on two things; Firstly, the magnitude of the difference between the empirical interval of* $|\beta_1|$ *and that of* $|\beta_n|$ *and secondly, the weights of the outcast samples, e.g. the samples of* $|\beta_1|$ *which lie outside the empirical interval for* $|\beta_n|$. *Therefore, if we consider Scaled GOE matrices, the magnitude of the difference between the empirical intervals for* $|\beta_1|$ *and* $|\beta_n|$ *will be smaller, compared to the corresponding difference in the case of GOE matrices. This is since the former difference will be equal to the latter scaled by a factor of* $\sqrt{\frac{2}{n}}$. *Also,*

	$n = 10$	$n = 20$	$n = 100$	$n = 200$	$n = 500$
GOE	0.0165	0.0142	0.0081	0.0075	0.0093 (!)
Sc. GOE	0.0161	0.0151	0.0078	0.0075	0.0053
Uniform	0.0365	0.0237	0.0097	0.0079	0.0075
Bernoulli	0.0651	0.0394	0.0133	0.0114	0.0062
Laplacian	0.07	0.0571	0.0254	0.0180	0.0069

Table 2.2: *The values of* $\max_t \left|G_n^{(S)}(t) - F_n^{(S)}(t)^2\right|$ *for all* 5 *types of matrices: GOE, Scaled GOE, Uniform, Bernoulli and Laplacian.*

if the number of simulations is small, the weight of the outcasts is larger and therefore the corresponding "distortion" in the distribution of $\|B\|_2$ *is higher. Hence, in order to overcome possible significant inconsistencies between theory and numerics, one needs to increase the number of simulations before one can rely on experimental results for some types of random matrices. This observation once again proves the advantages of numerical approximations of the c.d.f.'s of* β_n *and* $\|B\|_2$*, for example, via solutions of the Painlevé II initial value problem.*

The results from Corollaries 2.4.1 and 2.4.3 below are used in the next chapter. Corollary 2.4.1 uses Theorem 2.4.1 and Lemma 2.4.1 - results that are rigorously proved. We still don't know the proof of Corollary 2.4.3, but suspect that it is also linked with Conjecture 2.4.1.

Lemma 2.4.1. *Let* B *be a GOE matrix and* β_n *be its largest eigenvalue. Then*

$$\sqrt{\frac{2}{n}}\beta_n \xrightarrow{\mathcal{D}} 2. \tag{2.26}$$

Proof. From (2.4) we have

$$\zeta_n = n^{1/6}(\sqrt{2}\beta_n - 2\sqrt{n}) \qquad \text{and} \qquad \zeta_n \xrightarrow{\mathcal{D}} \zeta,$$

where ζ has a stationary distribution. Therefore

$$\frac{1}{n^{2/3}}\zeta_n = \sqrt{\frac{2}{n}}\beta_n - 2$$

and thus, using Slutsky's Theorem (c.f. Theorem A.1.7), we obtain

$$\frac{1}{n^{2/3}}\zeta_n \xrightarrow{\mathcal{D}} 0 \cdot \zeta = 0. \tag{2.27}$$

Hence

$$\sqrt{\frac{2}{n}}\beta_n - 2 \xrightarrow{\mathcal{D}} 0,$$

which is equivalent to (2.26) (by Slutsky's Theorem). □

Remark 2.4.8. *In the proof of Lemma 2.4.1, in order to obtain the right hand side of* (2.27)*, we have used the fact that* $\mathbb{P}[|\zeta| = \infty] = 0$.

Corollary 2.4.1. *Let* B *be a GOE matrix. Then, using Lemma 2.4.1 and Theorem 2.4.1, we have*

$$\sqrt{\frac{2}{n}}\|B\|_2 \xrightarrow{\mathcal{D}} 2,$$

as $n \to \infty$.

The proof is straightforward and is omitted.

Conjecture 2.4.3. *Let* B *be a GOE matrix. Then, assuming Conjecture 2.3.1, we have*

$$\sqrt{\frac{2}{n}}\,\mathbb{E}[\|B\|_2] \to 2,$$

as $n \to \infty$.

2.5. The invariance of GOE matrices with respect to multiplication by orthogonal matrices

In this section we prove that if $B \in \mathbb{R}^{n\times n}$ belongs to the Gaussian Orthogonal Ensemble, or shortly, if B is GOE matrix, and $V \in \mathbb{R}^{n\times n}$ is a deterministic orthogonal matrix, then the matrix $\tilde{B} := V^T B V$ is also GOE matrix. Also, in Corollary 2.5.1 we prove that the same invariance, with respect to multiplication by orthogonal matrices, holds for multiples of GOE matrices. (GOE matrices were defined in Definition 2.2.1.)

With the following lemma we prove that the entries of $\tilde{B}$ are normally distributed random variables with zero mean.

Lemma 2.5.1. *Let* $B \in \mathbb{R}^{n\times n}$ *be GOE matrix and* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$ *be deterministic vectors (not necessarily different from each other). Then the random variable* $\mathbf{u}^T B \mathbf{v}$ *is normally distributed and* $\mathbb{E}[\mathbf{u}^T B \mathbf{v}] = 0$.

Proof. Let us first expand $\mathbf{u}^T B \mathbf{v}$. We have

$$\mathbf{u}^T B \mathbf{v} = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j B_{ij} \tag{2.28}$$

and therefore the sum above can be rearranged in such a way that it becomes a sum of *independent* normally distributed random variables. Thus, by Theorem A.1.4, $\mathbf{u}^T B\mathbf{v}$ is a normally distributed random variable. Further,

$$\mathbb{E}[\mathbf{u^T}B\mathbf{v}] = \mathbb{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j B_{ij}\Big] = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j \mathbb{E}[B_{ij}] = 0. \qquad \square$$

In the following lemma, we use Definitions A.1.12 and A.1.13.

Lemma 2.5.2. *Let $B \in \mathbb{R}^{n\times n}$ be a GOE matrix and $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{z} \in \mathbb{R}^{n\times 1}$ be deterministic vectors of unit length (not necessarily different from each other). Then the random variables, $\mathbf{u}^T B\mathbf{v}$ and $\mathbf{w}^T B\mathbf{z}$, are jointly normally distributed.*

Proof. In order to prove that $\mathbf{u}^T B\mathbf{v}$ and $\mathbf{w}^T B\mathbf{z}$ are jointly normally distributed, according to Definitions A.1.12 and A.1.13, we have to prove that $c_1\mathbf{u}^T B\mathbf{v} + c_2\mathbf{w}^T B\mathbf{z}$ is a normally distributed random variable for any pair of constants $c_1, c_2 \in \mathbb{R}$. After expanding, as in (2.28), we obtain

$$c_1\mathbf{u}^T B\mathbf{v} + c_2\mathbf{w}^T B\mathbf{z} = \sum_{i=1}^{n}\sum_{j=1}^{n}(c_1 u_i v_j + c_2 w_i z_j)B_{ij},$$

which can be rearranged as a sum of *independent* normally distributed random variables. Hence, using Theorem A.1.4, we can infer that $c_1\mathbf{u}^T B\mathbf{v} + c_2\mathbf{w}^T B\mathbf{z}$ is a normally distributed random variable. $\square$

Remark 2.5.1. *It is clear from the proof of Lemma 2.5.2 that it can be extended in the following way: If $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ are any deterministic $n \times 1$ vectors, then the random variables $\mathbf{u}_1^T B\mathbf{v}_1, \mathbf{u}_2^T B\mathbf{v}_2, \ldots, \mathbf{u}_m^T B\mathbf{v}_m$ are jointly normally distributed, or equivalently, the vector $(\mathbf{u}_1^T B\mathbf{v}_1, \mathbf{u}_2^T B\mathbf{v}_2, \ldots, \mathbf{u}_m^T B\mathbf{v}_m)^T$ has an m-variate normal distribution.*

In fact, we shall use exactly this extension of Lemma 2.5.2 in Theorem 2.5.1. It will imply there that the entries above and on the main diagonal of the matrix $\tilde{B} = V^T BV$ are jointly normally distributed, which, combined with the fact that these entries are also uncorrelated (proved in Lemma 2.5.3), will lead to the conclusion that they are independent.

Lemma 2.5.3. *Let $B \in \mathbb{R}^{n\times n}$ be a GOE matrix and $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{z} \in \mathbb{R}^{n\times 1}$ be deterministic vectors of unit length (not necessarily different from each other). We calculate the value of $\mathbb{E}[(\mathbf{u}^T B\mathbf{v})(\mathbf{w}^T B\mathbf{z})]$ in the following cases:*

(a) If $\mathbf{u} = \mathbf{v} = \mathbf{w} = \mathbf{z}$, then $\mathbb{E}[(\mathbf{u}^T B\mathbf{u})^2] = 1$;

(b) If $\mathbf{u} = \mathbf{w}$, $\mathbf{v} = \mathbf{z}$ *and* $\mathbf{u} \perp \mathbf{v}$, *then* $\mathbb{E}[(\mathbf{u}^T B\mathbf{v})^2] = \frac{1}{2}$;

(c) If $(\mathbf{u}^T\mathbf{w})(\mathbf{v}^T\mathbf{z}) = 0$ *and* $(\mathbf{u}^T\mathbf{z})(\mathbf{v}^T\mathbf{w}) = 0$, *then* $\mathbb{E}[(\mathbf{u}^T B\mathbf{v})(\mathbf{w}^T B\mathbf{z})] = 0$;

Proof. Without assuming anything about the vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and $\mathbf{z}$ we obtain

$$\mathbb{E}[(\mathbf{u}^T B\mathbf{v})(\mathbf{w}^T B\mathbf{z})] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} u_i v_j w_k z_l \mathbb{E}[B_{ij}B_{kl}].$$

Therefore

$$\begin{aligned}\mathbb{E}[(\mathbf{u}^T B\mathbf{v})(\mathbf{w}^T B\mathbf{z})] &= \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j w_i z_j \mathbb{E}[B_{ij}^2] + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j w_j z_i \mathbb{E}[B_{ij}^2] \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j w_i z_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j w_j z_i, \end{aligned} \tag{2.29}$$

where we have used that $\mathbb{E}[B_{ii}^2] = 2\mathbb{E}[B_{ij}^2]$ when $i \neq j$ (see Definition 2.2.1). Hence, depending on the relations between $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and $\mathbf{z}$ in (2.29), we obtain

(a) If $\mathbf{u} = \mathbf{v} = \mathbf{w} = \mathbf{z}$, then

$$\mathbb{E}[(\mathbf{u}^T B\mathbf{u})^2] = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i^2 u_j^2 = 1, \qquad \text{since } \|\mathbf{u}\|_2 = 1;$$

(b) If $\mathbf{u} = \mathbf{w}$, $\mathbf{v} = \mathbf{z}$ and $\mathbf{u} \perp \mathbf{v}$, then

$$\mathbb{E}[(\mathbf{u}^T B\mathbf{v})^2] = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i^2 v_j^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j v_i u_j = \frac{1}{2}, \text{ since } \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1 \text{ and } \mathbf{u} \perp \mathbf{v};$$

(c) Let

$$(\mathbf{u}^T\mathbf{w})(\mathbf{v}^T\mathbf{z}) = 0 \qquad \text{and} \qquad (\mathbf{u}^T\mathbf{z})(\mathbf{v}^\mathbf{T}\mathbf{w}) = 0. \tag{2.30}$$

Then from (2.29) we obtain

$$\mathbb{E}[(\mathbf{u}^T B\mathbf{v})(\mathbf{w}^T B\mathbf{z})] = \frac{1}{2}(\mathbf{u}^T\mathbf{w})(\mathbf{v}^T\mathbf{z}) + \frac{1}{2}(\mathbf{u}^T\mathbf{z})(\mathbf{v}^\mathbf{T}\mathbf{w}) = 0.$$

The condition expressed by (2.30) can be split into the following four conditions: $\mathbf{u} \perp \mathbf{w}$ and $\mathbf{u} \perp \mathbf{z}$, or $\mathbf{u} \perp \mathbf{w}$ and $\mathbf{v} \perp \mathbf{w}$, or $\mathbf{v} \perp \mathbf{z}$ and $\mathbf{u} \perp \mathbf{z}$, or $\mathbf{v} \perp \mathbf{z}$ and $\mathbf{v} \perp \mathbf{w}$.

□

Theorem 2.5.1. *Let* $B \in \mathbb{R}^{n\times n}$ *be a GOE matrix and* $V \in \mathbb{R}^{n\times n}$ *be an orthogonal matrix. Then the matrix* $\tilde{B} := VBV^T$ *is also a GOE matrix.*

Proof. Let the rows of the matrix V be the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^{n\times 1}$. Then $\tilde{B}_{ij} = \mathbf{v}_i^T B \mathbf{v}_j$. From Lemma 2.5.1 it follows that the entries of the matrix $\tilde{B}$, $\tilde{B}_{ij}$, are normally distributed random variables with mean zero. From the fact that B is a symmetric matrix, we have that $\tilde{B}$ is also a symmetric matrix.

From Lemma 2.5.3, parts (a) and (b), we obtain $\mathrm{Var}[\tilde{B}_{ii}] = \mathbb{E}[\tilde{B}_{ii}^2] = 1$ and $\mathrm{Var}[\tilde{B}_{ij}] = \mathbb{E}[\tilde{B}_{ij}^2] = \frac{1}{2}$ for $i \neq j$, respectively.

From Lemma 2.5.2 we have that the elements of $\tilde{B}$ above and on the main diagonal are jointly normally distributed. Further, from Lemma 2.5.3, part (c), we have that they are also pairwise uncorrelated and thus pairwise independent. Combining these two facts together we finally obtain that the elements of $\tilde{B}$ above and on the main diagonal are independent, normally distributed random variables. □

Corollary 2.5.1. *Let V be an orthogonal matrix and $B_c \in \mathbb{R}^{n\times n}$ be a multiple of a GOE matrix, that is*

$$B_c := cB,$$

where B is a GOE matrix and $c \in \mathbb{R}$ is a constant. Then VB_cV^T is also a multiple of a GOE matrix.

Proof. Clearly $VB_cV^T = cVBV^T$ and by Theorem 2.5.1 VBV^T is a GOE matrix. □

2.6. Asymptotic results about the impact of the diagonal elements of SGOE matrices on the distribution of their norms

In §2.3 we found a way to approximate the c.d.f and the p.d.f. of $\|B\|_2$, where B was GOE matrix of size n. Using Conjecture 2.3.1, we showed that existing theory for the distribution of the largest eigenvalue of B can be extended, so that the distribution of $\|B\|_2$ is also approximated by solving the Painlevé II ODE (see §2.2 and §2.3 for details). Further, with the help of Conjecture 2.4.1, and in particular that of Theorem 2.4.1, we saw that the 2-norm of $\sqrt{\frac{2}{n}}B$, where B is a GOE matrix, can also be approximated using the solution of the same Painleveé II ODE. In §2.4 we called the matrices, which were equal to $\sqrt{\frac{2}{n}}B$, where B is GOE matrix, Scaled GOE matrices.

Before we explain the goal of this section, let us give the following definitions. Let $B \in \mathbb{R}^{n\times n}$ be a GOE matrix and c be a real number, independent of n. We consider the class of matrices

$$B_c := \sqrt{\frac{2}{n}}B + cD,$$

where D is a diagonal matrix whose diagonal entries, $D_{11}, D_{22}, \ldots, D_{nn}$, are i.i.d. random variables, distributed $\mathcal{N}\left(0, \frac{1}{n}\right)$. Also, the diagonal entries of D may be independent of the entries of B, or may coincide with the diagonal entries of B, depending on the situation we are in. In this section we show that

$$\left| \|B_c\|_2 - \sqrt{\frac{2}{n}}\|B\|_2 \right| \xrightarrow{\mathcal{D}} 0, \tag{2.31}$$

as $n \to \infty$. We prove (2.31) by showing that the r.v. $\eta_n := \max\{|\xi_1|, |\xi_2|, \ldots, |\xi_n|\}$, where ξ_i, $1 \leq i \leq n$, are i.i.d. random variables distributed $\mathcal{N}\left(0, \frac{1}{n}\right)$, converges to zero in distribution (see Theorem 2.6.2). We also prove that the rate of convergence of η_n to zero is $n^{-1/2+\varepsilon}$, where $\varepsilon > 0$ (see Corollary 2.6.1), and give numerical results which support that theory (see Experiment 2.6.1). The use of (2.31) is that it broadens the class of matrices whose 2-norms can be approximated using the solution to the Painlevé II ODE considered in §2.2 and §2.3.

As we mentioned above, our main goal in this section is to prove that $\|D\|_2 \xrightarrow{\mathcal{D}} 0$ as $n \to \infty$ (see Theorem 2.6.2) and also that the rate of that convergence is faster than $n^{-1/2+\varepsilon}$ for any $\varepsilon > 0$ and slower than $n^{-1/2}$ (see Corollary 2.6.1). Before we present the proofs, let us see their implication of these results.

Using the triangle inequality we have

$$\|B_c\| = \|\sqrt{\frac{2}{n}}B + cD\| \leq \sqrt{\frac{2}{n}}\|B\| + |c|\|D\|$$

and also

$$\sqrt{\frac{2}{n}}\|B\| = \|B_c - cD\| \leq \|B_c\| + |c|\|D\|$$

and therefore

$$\left| \sqrt{\frac{2}{n}}\|B\|_2 - \|B_c\|_2 \right| \leq |c|\|D\|_2 \to 0 \qquad \text{as } n \to \infty \tag{2.32}$$

faster than $n^{-1/2+\varepsilon}$ for any $\varepsilon > 0$, but slower than $n^{-1/2}$.

For the purpose of our work we will be mainly interested in the following two cases:

1. When $c = -\sqrt{2}$ and $D_{ii} = B_{ii}$, $1 \leq i \leq n$. This gives us zeroes on the diagonal of the matrix B_c and models random symmetric adjacency matrices;

2. When $c = -1$ and $D_{ii} = B_{ii}$, $1 \leq i \leq n$. This makes all the entries of B_c, including its diagonal entries, independent and identically distributed random variables.

The discussion that follows proves that $\|D\|_2 \xrightarrow{\mathcal{D}} 0$ as $n \to \infty$ and also gives the rate

of that convergence.

Lemma 2.6.1. *Let* $\eta_n := \max\{|\xi_1|, |\xi_2|, \ldots, |\xi_n|\}$*, where* $\xi_i \in \mathcal{N}(0,1)$*,* $1 \leq i \leq n$*, are independent and identically distributed (i.i.d.) random variables. Then* $\eta_n \xrightarrow{\mathcal{D}} \infty$*.*

Proof. We have

$$\{\eta_n < x\} = \bigcap_{i=1}^{n} \{|\xi_i| < x\}$$

and since $\xi_1, \xi_2, \ldots, \xi_n$ are independent (see Definition A.1.5),

$$\mathbb{P}[\eta_n < x] = \prod_{i=1}^{n} \mathbb{P}[|\xi_i| < x] = (\mathbb{P}[|\xi_1| < x])^n \to 0 \qquad \text{as } n \to \infty. \qquad \square$$

Let us now consider the sequence $\{\sigma(n)\eta_n\}_{n\in\mathbb{N}}$, where the function $\sigma : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is smooth enough and satisfies $\sigma(n) \to 0$ when $n \to \infty$. An example of a function σ, which satisfies those conditions, is $\sigma(n) = n^{-\beta}$ for $\beta > 0$.

Lemma 2.6.2. *Let* $\xi \in \mathcal{N}(0,1)$ *and consider the sequence* $\{\sigma(n)|\xi|\}_{n\in\mathbb{N}}$*. We have*

$$\sigma(n)|\xi| \xrightarrow{\mathcal{D}} 0, \qquad \text{as } n \to \infty.$$

Proof. Consider

$$\mathbb{P}[\sigma(n)|\xi| < x] = \mathbb{P}\left[-\frac{x}{\sigma(n)} < \xi < \frac{x}{\sigma(n)}\right] = 1 - 2\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right),$$

where $\mathbb{P}[\xi < y] = F(y) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{y} \exp\left(-\frac{t^2}{2}\right) dt$. Thus, it is easy to see that $F\left(\dfrac{x}{\sigma(n)}\right) \to 1$, as $n \to \infty$. This implies $\mathbb{P}[\sigma(n)|\xi| < x] \to 1$, as $n \to \infty$. $\square$

Therefore, for the sequence $\{\sigma(n)\eta_n\}_{n\in\mathbb{N}}$ we have

$$\begin{aligned} \mathbb{P}[\sigma(n)\eta_n < x] = \mathbb{P}\left[\eta_n < \frac{x}{\sigma(n)}\right] &= \mathbb{P}\left[-\frac{x}{\sigma(n)} < \xi_1 < \frac{x}{\sigma(n)}\right]^n \\ &= \left[1 - 2\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right)\right]^n \end{aligned} \tag{2.33}$$

and hence the limit $\lim_{n\to\infty} \mathbb{P}[\sigma(n)\eta_n < x]$ can not be found directly, as it is of the form $[1^\infty]$. In order to find that limit, we use Theorem 2.6.1, which is only L'Hospital's rule, and Lemma 2.6.3, which uses that rule for a particular sequence.

Theorem 2.6.1 (L'Hospital's Rule)**.** *Let* lim *stand for the limit* $\lim_{x\to c}$*,* $\lim_{x\to c^-}$*,* $\lim_{x\to c^+}$*,* $\lim_{x\to\infty}$*, or* $\lim_{x\to-\infty}$ *and suppose that* $\lim f(x)$ *and* $\lim g(x)$ *are both zero or*

are both $\pm\infty$*. If*

$$\lim \frac{f'(x)}{g'(x)}$$

has a finite value or if the limit is $\pm\infty$*, then*

$$\lim \frac{f(x)}{g(x)} = \lim \frac{f'(x)}{g'(x)}.$$

The following lemma is a standard result.

Lemma 2.6.3. *Let the sequence* $\{a_n\}_{n\in\mathbb{N}}$ *be such that* $a_n \to 0$*, as* $n \to \infty$ *and* $na_n \to l$*, as* $n \to \infty$*. Then* $(1 + a_n)^n \to e^l$*, as* $n \to \infty$*.*

Proof. We shall prove a more general result: If $f : \mathbb{R} \to \mathbb{R}$ is a function, satisfying $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} \frac{f(x)}{x} = l$, then $\lim_{x\to 0}(1 + f(x))^{\frac{1}{x}} = e^l$. We have

$$\ln(1 + f(x))^{\frac{1}{x}} = \frac{\ln(1 + f(x))}{x} = \frac{\ln(1 + f(x)) - \ln(1)}{f(x)} \cdot \frac{f(x)}{x}$$

and taking $\lim_{x\to 0}$ we get

$$\lim_{x\to 0} \ln(1 + f(x))^{\frac{1}{x}} = \lim_{x\to 0} \frac{\ln(1 + f(x)) - \ln(1)}{f(x)} \lim_{x\to 0} \frac{f(x)}{x} = l$$

and finally using the continuity of $\ln(\cdot)$ we obtain the desired result. □

It is now easy to see how we can apply Lemma 2.6.3 to the problem of finding $\lim_{n\to\infty} \mathbb{P}[\sigma(n)\eta_n < x]$. In (2.33), if we denote $a_n := -2\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right)$, we have $a_n \to 0$, as $n \to \infty$. Therefore, in order to apply Lemma 2.6.3, we have to find the limit $\lim_{n\to\infty} na_n$, or equivalently, find $\lim_{n\to\infty} -2n\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right)$. Instead, we shall find $\lim_{n\to\infty} n\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right)$ and then we shall multiply the result by -2.

Expanding $\lim_{n\to\infty} n\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right)$ we obtain

$$\lim_{n\to\infty} n\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right) = \lim_{n\to\infty} n \int_{\frac{x}{\sigma(n)}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

The expression on the right hand side can be rewritten as

$$\frac{\int_{\frac{x}{\sigma(n)}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt}{\frac{1}{n}},$$

where both, the numerator and the denominator, converge to 0 as $n \to \infty$. Therefore we can apply the L'Hospital's Rule (Theorem 2.6.1), but we will do that on a more

general expression,

$$\frac{\int_{\frac{x}{\sigma(1/\alpha)}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt}{\alpha},$$

where $\alpha \to 0^+$. The derivative of the denominator is equal to 1. Therefore

$$\begin{aligned}\lim_{\alpha\to 0^+} \frac{\int_{\frac{x}{\sigma(1/\alpha)}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt}{\alpha} &= \lim_{\alpha\to 0^+} \frac{d}{d\alpha}\left(\int_{\frac{x}{\sigma(1/\alpha)}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt\right) \\ &= \lim_{\alpha\to 0^+} \exp\left(-\frac{x^2}{2\sigma(1/\alpha)^2}\right) \frac{x\sigma'(1/\alpha)}{\alpha^2\sigma(1/\alpha)^2}.\end{aligned}$$

Hence, according to L'Hospital's Rule (Theorem 2.6.1),

$$\lim_{n\to\infty} n\left(1 - F\left(\frac{x}{\sigma(n)}\right)\right) = \lim_{\alpha\to 0^+} \exp\left(-\frac{x^2}{2\sigma(1/\alpha)^2}\right) \frac{x\sigma'(1/\alpha)}{\alpha^2\sigma(1/\alpha)^2}$$

if the limit on the right exists, or is equal to $+\infty$, or $-\infty$. Let us recall that the function $\sigma : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is smooth enough (so that the differentiation above is valid) and $\sigma(x) \to 0$, as $x \to \infty$. Therefore $\sigma(1/\alpha) \to 0$, as $\alpha \to 0^+$ and hence, loosely speaking, if $\sigma'(1/\alpha)$ is bounded, or doesn't converge to $+\infty$ or $-\infty$ "too fast", the limit on the right hand side exists and is equal to zero. However, in order to be more specific, we choose $\sigma(n) := n^{-\beta}$, where $\beta > 0$, and state that result as a theorem.

Theorem 2.6.2. *Let $n \in \mathbb{N}$ and $\eta_n = \max\{|\xi_1|, |\xi_2|, \ldots, |\xi_n|\}$, where $\xi_i \in \mathcal{N}(0,1)$ are i.i.d. random variables. Then, if $\beta > 0$, we have*

$$n^{-\beta}\eta_n \xrightarrow{\mathcal{D}} 0 \qquad \text{as } n \to \infty.$$

Proof. To prove this theorem, we have to show that

$$\lim_{\alpha\to 0^+} \exp\left(-\frac{x^2}{2\sigma(1/\alpha)^2}\right) \frac{x\sigma'(1/\alpha)}{\alpha^2\sigma(1/\alpha)^2} = 0 \tag{2.34}$$

for $\sigma(n) = n^{-\beta}$. Once we have done that, when we return to (2.33), this will then imply that

$$\lim_{n\to\infty} \mathbb{P}[n^{-\beta}\eta_n < x] = \lim_{n\to\infty} \left[1 - 2\left(1 - F(xn^{\beta})\right)\right]^n = e^0 = 1,$$

where the second equality is the application of Lemma 2.6.3. The last result says that for every $x > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}[n^{-\beta}\eta_n < x] = 1,$$

which means that $n^{-\beta}\eta_n \xrightarrow{\mathcal{D}} 0$, as $n \to \infty$, as required.

Finally, to show (2.34) we note that $\frac{d}{d\alpha}\alpha^{-\beta} = -\beta\alpha^{-\beta-1}$ and therefore (2.34) becomes

$$\lim_{\alpha\to 0^+} \exp\left(-\frac{x^2}{2\alpha^{2\beta}}\right)\frac{x\beta\alpha^{\beta+1}}{\alpha^2\alpha^{2\beta}} = x\lim_{\alpha\to 0^+} \exp\left(-\frac{x^2}{2\alpha^{2\beta}}\right)\frac{1}{\alpha^{\beta+1}} = 0. \qquad \square$$

From elementary Probability Theory we know that $n^{-\beta}\eta_n \xrightarrow{\mathcal{D}} 0$ implies $n^{-\beta}\eta_n \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$. We next state the result, which we will be using for the rate of convergence of $\eta_n/\sqrt{n}$ to zero.

Corollary 2.6.1. *In the settings of Theorem 2.6.2 we have*

$$n^{1/2-\varepsilon}\left(n^{-1/2}\eta_n\right) \xrightarrow{\mathcal{D}} 0 \qquad \text{as } n \to \infty$$

for any scalar $\varepsilon > 0$. In other words, $n^{-1/2}\eta_n = \mathcal{O}(n^{-1/2+\varepsilon})$.

This last result says that $\eta_n/\sqrt{n}$ converges to zero faster than $n^{-1/2+\varepsilon}$ for any $\varepsilon > 0$, but slower than $n^{-1/2}$.

Let us, as a final step, relate these results (Theorem 2.6.2 and Corollary 2.6.1) back to the problem of proving that $\|D\|_2 \xrightarrow{\mathcal{D}} 0$, as $n \to \infty$. Let us recall that, by definition, D is a diagonal matrix whose diagonal entries, $D_{11}, D_{22}, \ldots, D_{nn}$, are independent and identically distributed random variables with distribution $\mathcal{N}\left(0, \frac{1}{n}\right)$. By definition, we have

$$\begin{aligned} \|D\|_2 &= \max\{|D_{11}|, |D_{22}|, \ldots, |D_{nn}|\} \\ &= \frac{1}{\sqrt{n}}\max\{|n^{1/2}D_{11}|, |n^{1/2}D_{22}|, \ldots, |n^{1/2}D_{nn}|\}, \end{aligned}$$

where $n^{1/2}D_{ii} \in \mathcal{N}(0,1)$, $1 \le i \le n$. Therefore we can apply Theorem 2.6.2 and infer that $\|D\|_2 \xrightarrow{\mathcal{D}} 0$, as $n \to \infty$. Further, Corollary 2.6.1 gives the speed of that convergence.

We test the theory of this section in the following experiment.

Experiment 2.6.1. *In this experiment we test, by simulation, how tight inequality* (2.32) *is by comparing the c.d.f.'s of $|\|B\| - \|B_c\||$ and $\|cD\|$ (see the beginning of this section for definitions of B_c and D_c) for different values of n, $n = 100, 200, 500$ and* 1000*, where $c = -\sqrt{2}$. (We recall that this choice of c corresponds to B_c being an adjacency matrix.) We also test the speed of convergence of $\|cD\|$ to zero and compare it against the theoretical results of Corollary 2.6.1.*

We simulate 10 000 *samples of the matrix B, whose elements above and on the main*

diagonal are independent, identically distributed random variables satisfying

$$B_{ii} \in \mathcal{N}\left(0, \frac{2}{n}\right), \quad 1 \leq i \leq n, \qquad \textit{and} \qquad B_{ij} \in \mathcal{N}\left(0, \frac{1}{n}\right), \quad 1 \leq i < j \leq n.$$

For each sample of the matrix B we find the corresponding sample of $|\|B\| - \|B_c\||$ and that of $\|D_c\|$, and then, using the built-in MATLAB function `ecdf`*, we find the empirical c.d.f.'s of the random variables $|\|B\| - \|B_c\||$ and $\|D_c\|$. We also find*

$$M_{|\|B\|-\|B_c\||} := \max\{|\|B\| - \|B_c\|| \mid \textit{ over all samples of } B\}$$

and

$$M_{\|D_c\|} := \max\{\|D_c\| \mid \textit{ over all samples of } B\}$$

in order to obtain the rate of convergence of $|\|B\| - \|B_c\||$ and $\|D_c\|$ to zero, as $n \to \infty$. The values of $M_{|\|B\|-\|B_c\||}$ and $M_{\|D_c\|}$ for each considered value of n are given in the captions of Figures 2-24, 2-25, 2-26 and 2-27.

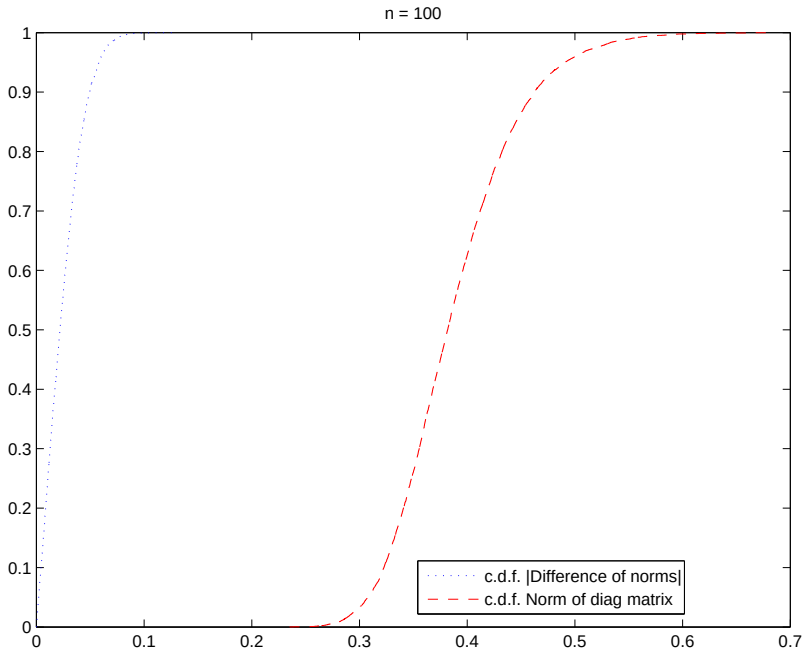

Figure 2-24: *Comparison between $|\|B\| - \|B_c\||$ and $\|cD\|$ when $n = 100$ and $c = -\sqrt{2}$, based on simulations (see (2.32)). $M_{\|cD\|} = 0.6819$ and $M_{|\|B\|-\|B_c\||} = 0.1287$.*

Results and Discussion. *From Figures 2-24, 2-25, 2-26 and 2-27 we can see that $\|D_c\|$ overestimates $|\|B\| - \|B_c\||$ by a factor of more than 2 for $n = 100$ and the difference between these two random variables increases as n increases. For example in*

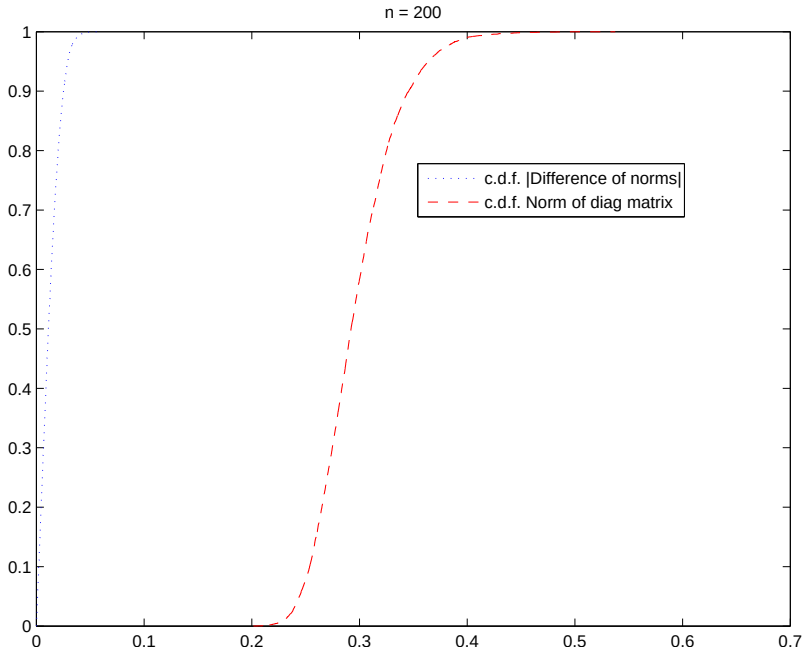

Figure 2-25: *Comparison between* $|\|B\|-\|B_c\|\|$ *and* $\|cD\|$ *when* $n = 200$ *and* $c = -\sqrt{2}$, *based on simulations (see* (2.32)*).* $M_{\|cD\|} = 0.5375$ *and* $M_{|\|B\|-\|B_c\|\|} = 0.0605$.

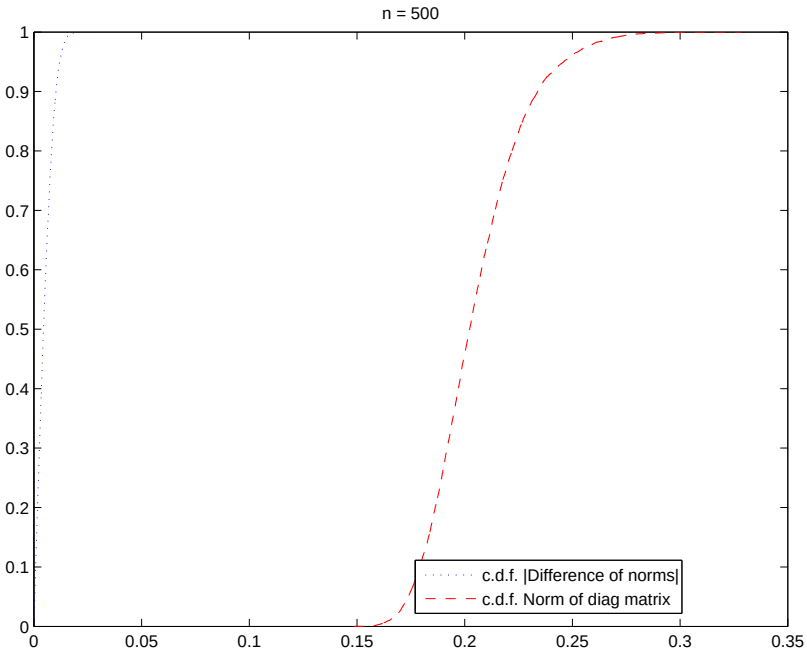

Figure 2-26: *Comparison between* $|\|B\|-\|B_c\|\|$ *and* $\|cD\|$ *when* $n = 500$ *and* $c = -\sqrt{2}$, *based on simulations (see* (2.32)*).* $M_{\|cD\|} = 0.3283$ *and* $M_{|\|B\|-\|B_c\|\|} = 0.0223$.

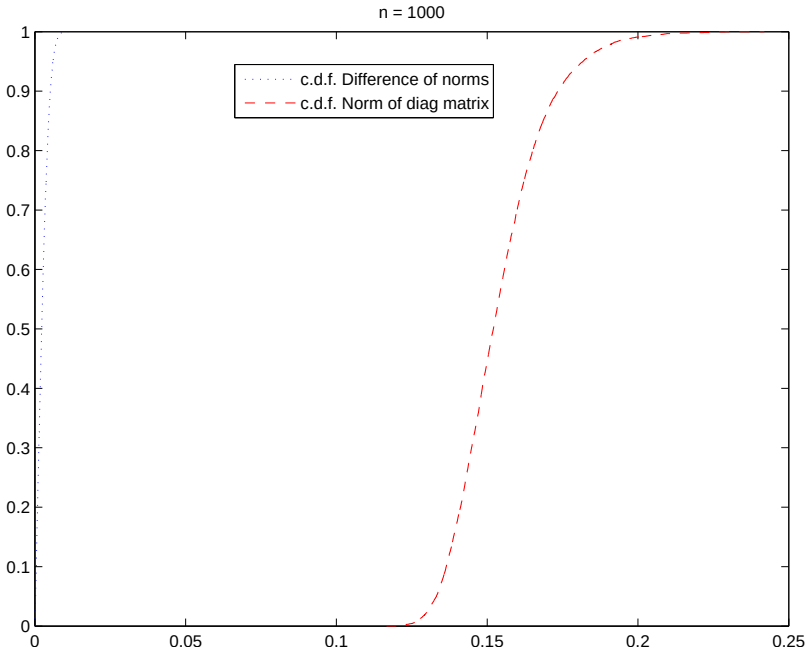

Figure 2-27: *Comparison between* $|\|B\| - \|B_c\||$ *and* $\|cD\|$ *when* $n = 1000$ *and* $c = -\sqrt{2}$*, based on simulations (see* (2.32)*).* $M_{\|cD\|} = 0.2419$ *and* $M_{|\|B\|-\|B_c\||} = 0.0110$.

Figure 2-24

$$\min\{\|D_c\| \mid \textit{ over all samples of } B\} \approx 0.25 \qquad \textit{and} \tag{2.35}$$
$$\max\{|\|B\| - \|B_c\|| \mid \textit{ over all samples of } B\} = 0.1287$$

and in Figures 2-26 and 2-27 the factor between the two quantities in (2.35) *is greater than* 3 *and* 5*, respectively. This means that inequality* (2.32) *is not sharp. Further, by considering the values of* $M_{|\|B\|-\|B_c\||}$ *for the different cases of* n*, we can see that* $M_{|\|B\|-\|B_c\||}$ *decreases similarly to* $\frac{C_1}{n}$ *for some constant* C_1*. By a similar inspection of the values of* $M_{\|D_c\|}$ *we can clearly see that they decrease similarly to* $\frac{C_2}{n^\alpha}$ *for some constants* C_2 *and* $\alpha < \frac{1}{2}$*. For example, calculating some of the values of* $M_{\|D_c\|}$ *we obtain:* $0.5375/0.6819 \approx 0.7882 > 1/\sqrt{2}$ *(for* $n = 100$ *and* 200*),* $0.2419/0.3283 \approx 0.7368 > 1/\sqrt{2}$ *(for* $n = 500$ *and* 1000*),* $0.2419/0.6819 \approx 0.3547 > 1/\sqrt{10} \approx 0.3162$ *(for* $n = 100$ *and* 1000*), which indicates a* $\frac{1}{\sqrt{n}}$ *dependance.*

Chapter 3. Stochastic versions of Weyl's Theorem

3.1. Introduction

In this chapter we shall consider perturbations of deterministic symmetric matrices by random symmetric matrices. Our aim is to extend known deterministic perturbation theory results to the stochastic case. In particular, we shall concentrate our attention on finding the largest possible magnitude of the perturbation, which in the same time doesn't cause a certain eigenvalue in the spectrum of the perturbed matrix to swap with the rest of the eigenvalues. However, the difference with deterministic perturbation theory is that when the perturbations are stochastic, we can only provide a so called *confidence level*, which in fact is a lower bound on the probability that there is no swap between our chosen eigenvalue and the rest of the perturbed spectrum.

The motivation for this work comes from considering the impact which sensitive data might have on the spectral clustering of networks. To explain this a little bit further, let us recall that in §1 we said that networks (or graphs) can be represented by their Laplacian matrix, or some other matrix associated with the network. These matrices usually contain the weights of the links (edges) between the vertices, but may contain other data related to the network. The idea of spectral clustering is to "extract" the important information from the matrix associated with the network by considering one or few of the leading eigenvectors of that matrix. Based on that, the network is clustered according to some criteria on the entries of these eigenvectors. Therefore, when the data contained in the matrix is sensitive to perturbations, one expects that this sensitivity is transferred, in a way, to the spectrum of the matrix and thus, to the spectral clustering of the network. Here we consider perturbations of deterministic matrices by random matrices, where the former can be thought of as the data matrix, obtained from a certain experiment, and the random perturbation to it could represent the correction of that data to the "real" data, which we have not been able to measure accurately, due to its sensitivity. So, the question which we address in this chapter

is: "What is the magnitude of the perturbation in the data, which would not lead to significant changes in the spectral clustering?" Here the term "significant changes in the spectral clustering" most of the time means that the eigenvalue, whose eigenvector is used for the clustering, swaps with other eigenvalue after the perturbation (see §3.3–§3.5). Also, the significance of the changes in the spectral clustering is measured by the magnitude of the angle between the eigenvector, which we use for the clustering, and its perturbed counterpart (see §3.7).

The background results which we use here include Weyl's inequality, the symmetric version of Bauer-Fike's Theorem and also two theorems about the gap in the spectrum of symmetric matrices (Theorems 3.2.2 and 3.2.3 below). All these theorems are first stated as they are in the deterministic case and are either adapted, or extended, to results about perturbations by random symmetric matrices.

In this chapter the *confidence level*, $1-\alpha$, where $0 < \alpha < 1$, will be given and our task will consist of finding a magnitude of perturbation, as large as possible, so that the probability that a given eigenvalue doesn't swap with any other eigenvalue in the spectrum, after the perturbation, is not less than $1-\alpha$. We now state this problem mathematically.

Problem 3.1.1. *Let $A \in \mathbb{R}^{n\times n}$ be a deterministic symmetric matrix and $B \in \mathbb{R}^{n\times n}$ be a random symmetric matrix. Further, let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ be the eigenvalues of A and B, respectively. Finally, let $\varepsilon \in \mathbb{R}$, $A(\varepsilon) := A + \varepsilon B$ and $\lambda_1(\varepsilon) \leq \lambda_2(\varepsilon) \leq \cdots \leq \lambda_n(\varepsilon)$ be the eigenvalues of $A(\varepsilon)$. Given an index $1 \leq k \leq n$, such that the eigenvalue λ_k is simple, and a* confidence level $1-\alpha$, *find an $\varepsilon^* > 0$, as large as possible, so that*

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon), \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon^* \textit{ and all } i \neq k] \geq 1-\alpha. \tag{3.1}$$

Ideally, we would like to find the largest possible magnitude ε^* which solves Problem 3.1.1, thst is, an ε^* for which

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon), \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon^* \text{ and all } i \neq k] = 1-\alpha.$$

In reality, we will not be able to find the exact distributions of $\lambda_i(\varepsilon)$, $1 \leq i \leq n$, and therefore we shall only use approximations of those distributions. This will in turn give us magnitudes ε^*, for which (3.1) will be a strict inequality, but our aim will be to approach the largest magnitude as much as we can.

In this chapter the random symmetric matrix B in Problem 3.1.1 will be assumed to belong to the class of Scaled GOE matrices, which we shall also call SGOE matrices for the sake of brevity. In §2.4 we defined what Scaled GOE matrix is, but we find it

convenient to define it again here.

Definition 3.1.1. We say that the random symmetric matrix B is Scaled GOE matrix, or shortly SGOE matrix, if its entries above and on the main diagonal are independent random variables satisfying

$$B_{ii} \in \mathcal{N}\left(0, \frac{2}{n}\right) \quad \text{for } 1 \leq i \leq n \qquad \text{and} \qquad B_{ij} \in \mathcal{N}\left(0, \frac{1}{n}\right) \quad \text{for } 1 \leq i < j \leq n. \tag{3.2}$$

We recall that if $B \in \mathbb{R}^{n\times n}$ is GOE matrix (c.f. §2) and $B^{(s)} \in \mathbb{R}^{n\times n}$ is SGOE matrix, the relation between $B^{(s)}$ and B is given by the equality

$$B^{(s)} = \sqrt{\frac{2}{n}} B.$$

Thus, $B^{(s)}$ is a *scaled* version of B.

The fact that we are mostly using SGOE matrices in this chapter, except in §3.8, means that we would mostly be relying on Theorem 2.4.1 and Corollary 2.4.1, which we have proved rigorously. We shall only be using Conjecture 2.4.3 when we derive (3.24) at the end of §3.4. In §3.8 we don't use any conjectures either.

Finally, we shall need the following notation.

Definition 3.1.2. Let $A \in \mathbb{R}^{n\times n}$ be symmetric matrix and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be its eigenvalues. Then we denote the gaps between the consecutive eigenvalues of A by γ_i, $1 \leq i \leq n-1$. More precisely,

$$\gamma_i := \lambda_{i+1} - \lambda_i, \quad 1 \leq i \leq n-1. \tag{3.3}$$

The plan of this chapter is as follows.

Firstly, in §3.2 we revise some well-known results about perturbations of the symmetric eigenvalue problem. Secondly, we apply Markov's inequality together with the Bauer-Fike Theorem to provide a first solution to Problem 3.1.1. Thirdly, we apply results from §2, where the probability density function (p.d.f.) and the cumulative density function (c.d.f.) of $\|B\|_2$ are computed to obtain an improved solution to Problem 3.1.1. Fourthly, we further improve the solution to Problem 3.1.1 by use of a finer perturbation result. These results are supported by numerical experiments in §3.9. Finally, we give a stochastic analogue of the perturbation of an eigenvector corresponding to a simple eigenvalue. This last result will have applications in spectral clustering of graphs, but we do not consider this aspect here.

3.2. Background Linear Algebra

Here we list some classic theorems in perturbation theory for symmetric eigenvalue problems, that we shall use in this chapter.

Theorem 3.2.1 (Weyl's inequality)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ be their eigenvalues, respectively. If $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$ are the eigenvalues of $A + B$, the following result holds:*

$$\lambda_i + \beta_1 \leq \tilde{\lambda}_i \leq \lambda_i + \beta_n \tag{3.4}$$

for all $1 \leq i \leq n$.

Proof. A proof of this result can be found in (Horn and Johnson, 1990), pp. 181–182, or in (Wilkinson, 1965), pp. 102–103. □

The following result is an analogue of the Bauer-Fike Theorem for symmetric matrices (c.f. (Wilkinson, 1965), pp. 102–103). In fact, the Bauer-Fike's Theorem holds for diagonalisable matrices, but its "symmetric version" can be obtained as a corollary from Weyl's inequality, Theorem 3.2.1. This is the reason for using the possibly misleading name "Bauer-Fike's Theorem" for the next corollary.

Corollary 3.2.1 (Bauer-Fike's Theorem)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and ε be a real number. The eigenvalues λ_i and β_i, $1 \leq i \leq n$, are defined as in Theorem 3.2.1. Let $\lambda_1(\varepsilon) \leq \lambda_2(\varepsilon) \leq \cdots \leq \lambda_n(\varepsilon)$ be the eigenvalues of $A + \varepsilon B$. Then*

$$\lambda_i - |\varepsilon|\|B\|_2 \leq \lambda_i(\varepsilon) \leq \lambda_i + |\varepsilon|\|B\|_2. \tag{3.5}$$

for all $1 \leq i \leq n$

Proof. The eigenvalues of B are either $\varepsilon\beta_1 \leq \varepsilon\beta_2 \leq \cdots \leq \varepsilon\beta_n$, when $\varepsilon \geq 0$, or $\varepsilon\beta_n \leq \varepsilon\beta_{n-1} \leq \cdots \leq \varepsilon\beta_1$, when $\varepsilon < 0$. Therefore, using Theorem 3.2.1 and the fact that

$$|\varepsilon|\|B\|_2 = \|\varepsilon B\|_2 = \max\{|\varepsilon\beta_1|, |\varepsilon\beta_n|\},$$

we obtain (3.5). □

Theorem 3.2.2. *((Parlett, 1998), Theorem 4.5.1) For any scalar σ and any nonzero vector $\mathbf{x}$ there is an eigenvalue λ of A satisfying*

$$|\lambda - \sigma| \leq \frac{\|A\mathbf{x} - \sigma\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

Proof. Without loss of generality we may assume that $\|\mathbf{x}\|_2 = 1$. If σ coincides with an eigenvalue of A, the result is immediate. Therefore, let us assume that $A - \sigma I$ is invertible. Then $\mathbf{x} = (A - \sigma I)^{-1}(A - \sigma I)\mathbf{x}$ and hence

$$1 = \|\mathbf{x}\|_2 \leq \|(A - \sigma I)^{-1}\|_2 \|(A - \sigma I)\mathbf{x}\|_2 = \frac{1}{\min_{1 \leq i \leq n} |\lambda_i - \sigma|} \|A\mathbf{x} - \sigma\mathbf{x}\|_2. \qquad \square$$

Theorem 3.2.3. *Let $A, B \in \mathbb{R}^{n \times n}$ be symmetric matrices, ε be a scalar and $A(\varepsilon) = A + \varepsilon B$. Let $\mathbf{y}$ be a unit vector, $\theta := \mathbf{y}^T(A + \varepsilon B)\mathbf{y}$ and the residual $r(\mathbf{y})$ be defined as $r(\mathbf{y}) := (A + \varepsilon B)\mathbf{y} - \theta\mathbf{y}$. Let $\lambda(\varepsilon)$ be the eigenvalue of $A + \varepsilon B$ closest to θ. Then*

$$|\theta - \lambda(\varepsilon)| \leq \frac{\|r(\mathbf{y})\|^2}{\gamma_\theta}, \tag{3.6}$$

where $\gamma_\theta := \min\{|\lambda_i(\varepsilon) - \theta| \mid 1 \leq i \leq n, \lambda_i(\varepsilon) \neq \lambda(\varepsilon)\}$.

Proof. A proof of this result can be found in (Parlett, 1998), pp. 244-246. $\square$

Next we consider the question of bounding $|\sin\psi|$, where $\psi = \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$ for some index $1 \leq k \leq n$. We start with a Theorem from (Parlett, 1998), Theorem 11.7.1 there. Let $A \in \mathbb{R}^{n \times n}$ is a deterministic symmetric matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, where the eigenvalue λ_k will be assumed simple. We shall also need the definition of the gap between a given number, say θ, and the spectrum of A. This definition was already given in the statement of Theorem 3.2.3, but we shall recall it here.

Let $\theta \in \mathbb{R}$ be a number and λ_k be the eigenvalue in the spectrum of A, which is closest to it, that is,

$$|\lambda_k - \theta| = \min_{1 \leq j \leq n} |\lambda_j - \theta|.$$

Then the gap between θ and the spectrum of A is defined as

$$\gamma_\theta := \min_{j, j \neq k} |\lambda_j - \theta|.$$

Theorem 3.2.4. *Let $\mathbf{y}$ be a unit vector and $\theta \in \mathbb{R}$ be some number. Let us also define the residual, $r(\mathbf{y}) := A\mathbf{y} - \mathbf{y}\theta$. Further, let λ_k be the eigenvalue of A closest to θ and $\mathbf{v}_k$ be its corresponding eigenvector of unit length. If $\psi := \angle(\mathbf{y}, \mathbf{v}_k)$, we have*

$$|\sin\psi| \leq \frac{\|r(\mathbf{y})\|_2}{\gamma_\theta}. \tag{3.7}$$

Proof. The proof can be found in (Parlett, 1998), pp. 244-246. $\square$

The following theorem provides a lower bound on $|\sin\psi|$.

Theorem 3.2.5. *Let* $\mathbf{y}$ *be a unit vector with* $\theta = \mathbf{y}^T A\mathbf{y}$ *and residual* $r(\mathbf{y}) = A\mathbf{y} - \mathbf{y}\theta$. *Let* α *be an eigenvalue of* A *and let* $\mathbf{z}$ *be its normalised eigenvector, and let* $\psi := \angle(\mathbf{y}, \mathbf{z})$. *Then*

$$|\sin\psi| \geq \frac{\|r(\mathbf{y})\|}{\text{spread}(A)}, \tag{3.8}$$

where $\text{spread}(A) := \max_{i,j} |\lambda_i - \lambda_j| = \lambda_n - \lambda_1$.

Proof. The proof can be found in (Parlett, 1998), pp. 244-246. □

Next we shall extend the results in this section to the case when the matrix by which we perturb, B, is SGOE matrix.

3.3. An extension to the Bauer-Fike Theorem using Markov's inequality

In this section we recall Markov's inequality, which is further combined with Bauer-Fike's Theorem and the result is an extension of the latter to the case when B is a random matrix.

Theorem 3.3.1 (Markov's inequality)**.** *Let* X *be a random variable and* a *be a positive number. Then*

$$\mathbb{P}[|X| \geq a] \leq \frac{\mathbb{E}[|X|]}{a}. \tag{3.9}$$

Proof. It is easy to see that

$$aI_{|X|\geq a} \leq |X|,$$

where $I_{|X|\geq a}$ is the indicator of the *event* $\{|X| \geq a\}$. Therefore we may take expectations on both sides of the last inequality and get

$$a\mathbb{E}[I_{|X|\geq a}] \leq \mathbb{E}[|X|],$$

since a is positive. Then (3.9) follows from the last inequality and the identity $\mathbb{E}[I_{|X|\geq a}] = \mathbb{P}[|X| \geq a]$. □

Remark 3.3.1. *The sharpness of the Markov's inequality depends on two things: on the choice of the number* a *and on the probability measure of the event* $\{X < a\}$. *For example, if* X *is a random variable on the interval* $[0, 1]$ *with the usual Lebesgue measure and if*

$$X(\omega) := \begin{cases} 1 & \text{if } \omega \geq 0.5; \\ 0 & \text{if } \omega < 0.5, \end{cases}$$

then choosing $a = 1$ *makes Markov's inequality sharp, but choosing* $a = 0.5$ *will produce* $0.25 = 0.5\mathbb{P}[|X(\omega)| \geq 0.5] < \mathbb{E}[|X(\omega)|] = 0.5$. *And if, as another example, we choose* $X(\omega)$ *to be a uniformly distributed random variable on the interval* $[0, 1]$, *we get*

$$a(1-a) = a\mathbb{P}[|X(\omega)| \geq a] < \mathbb{E}[|X(\omega)|] = 0.5.$$

Thus, choosing $a = 0.5$ *maximises the left hand side, making it equal to* 0.25, *but this is still only* 50% *of the right hand side. Thus, loosely speaking, if* X *is a random variable which is very similar to a step function with one of its steps clearly dominating the rest by its height, then Markov's inequality may be sharp, but otherwise it can be very crude.*

So far, in Theorem 3.2.1 and in Corollary 3.2.1, A and B were deterministic symmetric matrices. Let us assume now that the entries of B are random variables. This implies that the eigenvalues of B, $\beta_1, \beta_2, \ldots, \beta_n$, and those of $A + \varepsilon B$, $\lambda_1(\varepsilon), \lambda_2(\varepsilon), \ldots, \lambda_n(\varepsilon)$, are random variables and that inequalities (3.4) and (3.5) hold with probability one. The following result is an analogue of inequalities (3.4) and (3.5) in probabilistic form.

Theorem 3.3.2. *Let* $A \in \mathbb{R}^{n\times n}$ *be a deterministic symmetric matrix and* $B \in \mathbb{R}^{n\times n}$ *be a random symmetric matrix. Let also* $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ *and* $\lambda_1(\varepsilon) \leq \lambda_2(\varepsilon) \leq \cdots \leq \lambda_n(\varepsilon)$ *be the eigenvalues of* A *and* $A(\varepsilon) = A + \varepsilon B$, *respectively, where* $\varepsilon \in \mathbb{R}$. *Then the following inequality holds:*

$$\mathbb{P}\left[\exists i,\ s.t.\ |\lambda_i - \lambda_i(\varepsilon)| \geq \delta\right] \leq |\varepsilon| \frac{\mathbb{E}[\|B\|_2]}{\delta}, \tag{3.10}$$

where $\delta > 0$.

Proof. When the Bauer-Fike Theorem is interpreted probabilistically we obtain

$$\{\exists i,\ \text{s.t. } |\lambda_i - \lambda_i(\varepsilon)| \geq \delta\} \subset \{|\varepsilon|\|B\|_2 \geq \delta\},$$

where $\{\exists i,\ \text{s.t. } |\lambda_i - \lambda_i(\varepsilon)| \geq \delta\}$ and $\{|\varepsilon|\|B\|_2 \geq \delta\}$ denote the probabilistic events in the curly brackets, respectively. Therefore

$$\mathbb{P}[\exists i,\ \text{s.t. } |\lambda_i - \lambda_i(\varepsilon)| \geq \delta] \leq \mathbb{P}[|\varepsilon|\|B\|_2 \geq \delta]. \tag{3.11}$$

Further, from Markov's inequality we have,

$$\mathbb{P}\left[|\varepsilon|\|B\|_2 \geq \delta\right] \leq \frac{\mathbb{E}\left[|\varepsilon|\|B\|_2\right]}{\delta},$$

which, together with (3.11), gives us (3.10). □

Theorem 3.3.2 is an easy extension of Corollary 3.2.1 to the case when B is a general random symmetric matrix. Inequality (3.10) is useful when we know $\mathbb{E}[\|B\|_2]$, or can estimate it. Below we give two alternative ways of approximating $\mathbb{E}[\|B\|_2]$ in the case when B is SGOE matrix, but for the moment we shall present a way of calculating a solution to Problem 3.1.1, under the assumption that $\mathbb{E}[\|B\|_2]$ is accessible.

Theorem 3.3.3. *Let the* confidence level $1-\alpha$ *be given, where* $0<\alpha<1$*, and* δ *be some positive number. Then, in the notation of Theorem 3.3.2, if we choose*

$$\varepsilon_1^* := \frac{\alpha\delta}{\mathbb{E}[\|B\|_2]}, \tag{3.12}$$

we obtain

$$\mathbb{P}\left[|\lambda_i - \lambda_i(\varepsilon)| < \delta, \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \le \varepsilon_1^* \textit{ and all } 1 \le i \le n\right] \ge 1-\alpha.$$

Proof. From the Bauer-Fike Theorem we have

$$\{|\lambda_i - \lambda_i(\varepsilon)| < \delta \quad \text{for all } 1 \le i \le n\} \supset \{|\varepsilon|\|B\|_2 < \delta\}$$

for all ε. Thus,

$$\bigcap_{\varepsilon,|\varepsilon|\le\varepsilon_1^*} \{|\lambda_i - \lambda_i(\varepsilon)| < \delta \quad \text{for all } 1 \le i \le n\} \supset \bigcap_{\varepsilon,|\varepsilon|\le\varepsilon_1^*} \{|\varepsilon|\|B\|_2 < \delta\} = \{\varepsilon_1^*\|B\|_2 < \delta\} \tag{3.13}$$

and since

$$\begin{aligned} &\{|\lambda_i - \lambda_i(\varepsilon)| < \delta, \quad \text{for all } \varepsilon \text{ s.t. } |\varepsilon| \le \varepsilon_1^* \text{ and all } 1 \le i \le n\} \\ = &\bigcap_{\varepsilon,|\varepsilon|\le\varepsilon_1^*} \{|\lambda_i - \lambda_i(\varepsilon)| < \delta \quad \text{for all } 1 \le i \le n\}, \end{aligned}$$

after taking probabilities in (3.13) we obtain

$$\begin{aligned} \mathbb{P}\left[|\lambda_i - \lambda_i(\varepsilon)| < \delta, \quad \text{for all } \varepsilon \text{ s.t. } |\varepsilon| \le \varepsilon_1^* \text{ and all } 1 \le i \le n\right] &\ge \mathbb{P}[\varepsilon_1^*\|B\|_2 < \delta] \\ &\ge 1 - \varepsilon_1^* \frac{\mathbb{E}[\|B\|_2]}{\delta}. \end{aligned}$$

Thus, from the definition of ε_1^*, (3.12), we finally obtain

$$\mathbb{P}\left[|\lambda_i - \lambda_i(\varepsilon)| < \delta, \quad \text{for all } \varepsilon \text{ s.t. } |\varepsilon| \le \varepsilon_1^* \text{ and all } 1 \le i \le n\right] \ge 1-\alpha,$$

which completes the proof. □

The next corollary shows that when ε_1^* is given by (3.12), where $\delta := \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}$, then ε_1^* solves Problem 3.1.1.

Corollary 3.3.1. *In the notation of Theorems 3.3.2 and 3.3.3 let us assume that the eigenvalue λ_k of A is simple. Let us further define ε_1^* as in* (3.12)*, with δ replaced by $\frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}$, that is, let*

$$\varepsilon_1^* := \frac{\alpha \min\{\gamma_{k-1}, \gamma_k\}}{2\mathbb{E}[\|B\|_2]}.$$

Then we have

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \textit{for all } \varepsilon, \textit{ s.t. } |\varepsilon| \leq \varepsilon_1^* \textit{ and all } i \neq k] \geq 1 - \alpha,$$

where the confidence level*, $1 - \alpha$, is given.*

Proof. It is easy to show that when

$$|\lambda_i - \lambda_i(\varepsilon)| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\} \qquad \text{for all } 1 \leq i \leq n,$$

then

$$\lambda_i(\varepsilon) < \lambda_k(\varepsilon) \quad \text{for all } 1 \leq i < k \qquad \text{and} \qquad \lambda_k(\varepsilon) < \lambda_j(\varepsilon) \quad \text{for all } k < j \leq n,$$

which implies

$$\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \qquad \text{for all } i \neq k.$$

Therefore, in terms of (probabilistic) events we have

$$\begin{aligned} & \left\{ |\lambda_i - \lambda_i(\varepsilon)| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}, \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_1^* \text{ and all } 1 \leq i \leq n \right\} \\ \subset \; & \{\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_1^* \text{ and all } i \neq k\}. \end{aligned}$$

Hence, using Theorem 3.3.3, we obtain

$$\begin{aligned} 1 - \alpha \;&\leq\; \mathbb{P}\left[|\lambda_i - \lambda_i(\varepsilon)| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\} \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_1^* \text{ and all } 1 \leq i \leq n \right] \\ &\leq\; \mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_1^* \text{ and all } i \neq k] \end{aligned}$$

where

$$\varepsilon_1^* := \frac{\alpha \min\{\gamma_{k-1}, \gamma_k\}}{2\mathbb{E}\|B\|_2}. \qquad \square$$

Remark 3.3.2. *As it was already noted above, Theorems 3.3.2 and 3.3.3 and Corol-*

lary 3.3.1 hold for any random symmetric matrix, B. However, in order for these results to be applied in practice, one needs to be able to calculate $\mathbb{E}[\|B\|_2]$ for the particular random symmetric matrix B one considers. In §2 we showed how one could approximate the cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of $\|B\|_2$ numerically, when B is GOE or SGOE matrix. Below we extend Program 2.3.1 (given in §2) so that it also calculates $\mathbb{E}[\|B\|_2]$, when B is GOE or SGOE matrix.

We now discuss two ways of obtaining $\mathbb{E}[\|B\|_2]$ when B is SGOE matrix. Firstly, we could find $\mathbb{E}[\|B\|_2]$ by using known convergence results about $\|B\|_2$. In §2 we proved that $\|B\|_2 \xrightarrow{\mathcal{D}} 2$, as $n \to \infty$, when B is SGOE matrix. We also stated in Conjecture 2.4.3 (without a rigorous proof for the moment, but supported by experiments) that in that case

$$\mathbb{E}\left[\|B\|_2\right] \to 2, \qquad \text{as } n \to \infty.$$

Therefore, if we assume that B is SGOE matrix, we can replace $\mathbb{E}[\|B\|_2]$ with 2 in the definition of ε_1^*, that is, in (3.12).

Alternatively, we could find $\mathbb{E}[\|B\|_2]$ when B is GOE matrix, by using the results from §2.2, and then scale the value of $\mathbb{E}[\|B\|_2]$ by $\sqrt{\frac{2}{n}}$. In §2.2 we used a modified version of a MATLAB program by (Edelman and Persson, 2005), which calculates the distribution of $\|B\|_2$, where B is GOE matrix, as a numerical solution to the following system of Painlevé II ODEs:

$$\frac{d}{ds}\begin{pmatrix} q \\ q' \\ I \\ I' \\ J \end{pmatrix} = \begin{pmatrix} q' \\ sq + 2q^3 \\ I' \\ q^2 \\ -q \end{pmatrix} \tag{3.14}$$

with initial conditions

$$\begin{pmatrix} q(s_0) \\ q'(s_0) \\ I(s_0) \\ I'(s_0) \\ J(s_0) \end{pmatrix} = \begin{pmatrix} \mathrm{Ai}(s_0) \\ \mathrm{Ai}'(s_0) \\ \int_{s_0}^{\infty}(x - s_0)\mathrm{Ai}(x)^2 dx \\ \mathrm{Ai}(s_0)^2 \\ \int_{s_0}^{\infty}\mathrm{Ai}(x)dx \end{pmatrix}, \tag{3.15}$$

where s_0 is chosen large enough and the functions $I(s)$ and $J(s)$ were given by

$$I(s) := \int_s^{\infty}(x - s)q(x)^2 dx \qquad \text{and} \qquad J(s) := \int_s^{\infty} q(x)dx$$

(see §2.2 for more details). The functions

$$F(s) = \exp\left(-\frac{1}{2}(I(s) + J(s))\right) \qquad \text{and} \qquad G(s) = F(s)^2 \tag{3.16}$$

were the cumulative distribution functions (c.d.f.'s) of β_n, the largest eigenvalue of B, and $\|B\|_2$ with respect to the *edge scaling variable*, s.

Now we shall introduce $\mathbb{E}[\|B\|_2]$ as a new entry in the system of ODEs (3.14), with initial conditions (3.15). Let us define

$$E(s) := \int_s^\infty tg(t)dt,$$

where $g(t)$ is the probability density function (p.d.f.) of $\|B\|_2$, when B is GOE matrix. Then we have $\mathbb{E}[\|B\|_2] = E(-\infty)$. Therefore

$$\frac{d}{ds}E(s) = -sg(s),$$

with corresponding boundary condition $E(s) \to 0$ as $x \to \infty$. The latter will thus become the initial condition $E(s_0) = 0$ in the extended system of ODEs. Let us recall that, under the assumption $G(s) = F^2(s)$ for $s \in \mathbb{R}$ (see Conjecture 2.3.1), we have

$$g(s) = \frac{d}{ds}G(s) = \frac{d}{ds}F^2(s) = 2F(s)\frac{d}{ds}F(s) = 2F(s)f(s),$$

where $f(s)$ is the probability density function (p.d.f.) of β_n. Also, from (3.16) we have that

$$f(s) = \frac{d}{ds}F(s) = -\frac{1}{2}(I'(s) + J'(s))F(s) = -\frac{1}{2}(I'(s) + J'(s))\exp\left(-\frac{1}{2}(I(s) + J(s))\right).$$

Hence we obtain

$$\begin{aligned}\frac{d}{ds}E(s) = -sg(s) &= s(I'(s) + J'(s))\exp(-(I(s) + J(s))) \\ &= s(I'(s) - q(s))\exp(-(I(s) + J(s)))\end{aligned}$$

and thus the extended system of ODEs becomes

$$\frac{d}{ds}\begin{pmatrix} q \\ q' \\ I \\ I' \\ J \\ E \end{pmatrix} = \begin{pmatrix} q' \\ sq + 2q^3 \\ I' \\ q^2 \\ -q \\ s(I'(s) - q(s))\exp(-(I(s) + J(s))) \end{pmatrix} \tag{3.17}$$

with initial conditions

$$\begin{pmatrix} q(s_0) \\ q'(s_0) \\ I(s_0) \\ I'(s_0) \\ J(s_0) \\ E(s_0) \end{pmatrix} = \begin{pmatrix} \mathrm{Ai}(s_0) \\ \mathrm{Ai}'(s_0) \\ \int_{s_0}^{\infty}(x - s_0)\mathrm{Ai}(x)^2 dx \\ \mathrm{Ai}(s_0)^2 \\ \int_{s_0}^{\infty}\mathrm{Ai}(x)dx \\ 0 \end{pmatrix}. \tag{3.18}$$

One extra thing that we should take care of in this approach is the re-scaling from s, back to t. We recall that

$$s := n^{1/6}(\sqrt{2}t - 2\sqrt{n})$$

and therefore finding $\mathbb{E}[\|B\|_2]$ in terms of the *edge scaling variable*, s, is similar to finding $\mathbb{E}[n^{1/6}\left(\sqrt{2}\|B\|_2 - 2\sqrt{n}\right)]$, which we have to re-scale and shift back to $\mathbb{E}[\|B\|_2]$. This is in the case when B is GOE matrix. If B is a matrix defined by (3.2), that is, when B is SGOE matrix, we have to further multiply the result by $\sqrt{\frac{2}{n}}$. Hence, the result from the Painleve II system of ODEs, (3.17), with initial conditions (3.18), is in terms of the variable s and has to be divided by $n^{-2/3}$ and the shifted by 2, in order to get $\mathbb{E}[\|B\|_2]$ when B is SGOE matrix.

We now give the MATLAB program which solves the system of ODEs (3.17) with initial conditions (3.18). This program is a modification of the program given in (Edelman and Persson, 2005) which calculates $F(s)$, the c.d.f. of the largest eigenvalue of B, β_n, with respect to the *edge scaling variable*, s. Our program finds $\mathbb{E}[\|B\|_2]$ and $\mathrm{Var}[\|B\|_2]$ when B is GOE matrix. Therefore, when B is SGOE matrix, the outputs, `expect_norm` and `var_norm`, have to be multiplied further by $\sqrt{\frac{2}{n}}$.

Program 3.3.1. `% Defining n.`

```
n = 1e4;

% The right hand sides of the system of ODEs. We have added
```

```
% the right hand sides for the sixth and seventh equation,
% corresponding to the mean and the variance.
deq = inline('[y(2); s*y(1) + 2*y(1)^3; y(4); y(1)^2; -y(1); ...
       s*(y(4) - y(1))*exp(-y(5) - y(3)); ...
       (y(4)-y(1))*(exp(-y(5) - y(3)))*(s-y(6))^2]', ...
       's', 'y');

% The discrete interval over which we solve the ODE.
s0 = 5;
sn = -8;
sspan = linspace(s0, sn, 1000);

% Initial conditions.
y0 = [airy(s0); airy(1,s0);...
    quadl(inline('(x - s0).*airy(x).^2', 'x', 's0'), s0, 20, 1e-25, 0, s0); ...
    airy(s0)^2; quadl(inline('airy(x)'), s0, 20, 1e-18); 0; 0];

% Invoking the ODE solver.
opts = odeset('reltol', 1e-13, 'abstol', 1e-15);
[s, y] = ode45(deq, sspan, y0, opts);

% Finding the mean and the variance of the norm of B,
% when B is a GOE matrix.
expect_norm = (n^(-1/6)*interp1q(s, y(:, 6), sn) + 2*sqrt(n))/sqrt(2);
var_norm = (n^(-1/3))*interp1q(s, y(:, 7), sn)/2;
```

Therefore we now know a way of approximating $\mathbb{E}[\|B\|_2]$ numerically, when B is SGOE matrix and so we can find an ε^* in Problem 3.1.1, stated at the beginning of this chapter, by letting

$$\varepsilon_1^* := \frac{\alpha \min\{\gamma_{k-1}, \gamma_k\}}{2\mathbb{E}[\|B\|_2]}. \tag{3.19}$$

Numerical experiments, which test the result of Corollary 3.3.1, with ε_1^* given by (3.19), are given in §3.9 (see Experiment 3.9.1).

3.4. An extension to the Bauer-Fike Theorem using numerical approximations of the 2-norm of SGOE matrices

Now we start with the description of a slightly different approach for obtaining ε^* in Problem 3.1.1, which we call ε_2^* here. This new approach requires more information about the distribution of $\|B\|_2$, but it is similar to the method for obtaining ε_1^* in that it also uses the symmetric version of the Bauer-Fike Theorem. However, the difference between these two approaches is that the one we are going to present now doesn't use the Markov's inequality. We use this at the end to show that

$$\frac{\varepsilon_1^*}{\varepsilon_2^*} \approx \alpha,$$

where $1-\alpha$ is the *confidence level*, and hence $\varepsilon_2^* > \varepsilon_1^*$. We confirm this result experimentally in §3.9 (see Experiment 3.9.1).

In §2 we gave a way of finding the distribution of the 2-norm of GOE matrices (c.f. Definition 3.1.1). Using that we can obtain the distribution of $\|B\|_2$, when B is SGOE matrix, by scaling the distribution of the 2-norm of $n \times n$ GOE matrix by $\sqrt{\frac{2}{n}}$. However, the results which we are going to state now are valid for general random symmetric matrices.

Let $\hat{\varepsilon} > 0$ be some given number. As we saw earlier, from the Bauer-Fike Theorem we have:

$$|\lambda_i(\varepsilon) - \lambda_i| \leq |\varepsilon|\|B\|_2 \quad \text{for all } \varepsilon \text{ and all } 1 \leq i \leq n,$$

which interpreted probabilistically implies (see (3.13))

$$\mathbb{P}[|\lambda_i(\varepsilon) - \lambda_i| < \delta \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \hat{\varepsilon} \text{ and all } 1 \leq i \leq n] \geq \mathbb{P}[\hat{\varepsilon}\|B\|_2 < \delta]. \tag{3.20}$$

Further, we can show that

$$\begin{aligned} &\left\{|\lambda_i - \lambda_i(\varepsilon)| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\} \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \hat{\varepsilon} \text{ and all } 1 \leq i \leq n\right\} \\ \subset\ &\{\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \hat{\varepsilon} \text{ and all } i \neq k\}, \end{aligned} \tag{3.21}$$

like in the proof of Corollary 3.3.1. Therefore, after taking probabilities and using (3.20) with $\delta := \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}$, we obtain

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \hat{\varepsilon} \text{ and all } i \neq k] \geq \mathbb{P}\left[\hat{\varepsilon}\|B\|_2 < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}\right]. \tag{3.22}$$

Hence, if $G_n(t)$ is the cumulative distribution function (c.d.f.) of $\|B\|_2$ when $B \in \mathbb{R}^{n\times n}$, that is, if

$$G_n(t) := \mathbb{P}[\|B\|_2 < t],$$

then the following result holds.

Theorem 3.4.1. *Let the* confidence level, $1-\alpha$, *be given and* $t_2^* := G_n^{-1}(1-\alpha)$, *where* G_n^{-1} *is the inverse of the c.d.f.* G_n. *Then, in the notation of this chapter, if we let*

$$\varepsilon_2^* := \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2t_2^*}, \tag{3.23}$$

we have

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon_2^* \textit{ and all } i \neq k] \geq 1-\alpha.$$

Proof. Let ε_2^* be defined by (3.23) and ε be such that $|\varepsilon| \leq \varepsilon_2^*$ (and $\varepsilon \neq 0$). Then

$$G_n^{-1}(1-\alpha) = t_2^* = \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2\varepsilon_2^*}$$

and therefore, after applying G_n to both sides of the last equality, we obtain

$$1-\alpha = G_n\left(\frac{\min\{\gamma_{k-1}, \gamma_k\}}{2\varepsilon_2^*}\right) = \mathbb{P}\left[\|B\|_2 < \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2\varepsilon_2^*}\right] = \mathbb{P}\left[\varepsilon_2^*\|B\|_2 < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}\right].$$

Then the result follows from (3.22), after replacing the arbitrary positive $\hat{\varepsilon}$ there with ε_2^*. □

Remark 3.4.1. *Since Bauer-Fike's Theorem holds in a rather general setting, we expect* $\|B\|_2$ *to usually overestimate the difference* $|\lambda_i - \lambda_i(\varepsilon)|$ *by a rather large margin. Further, the inequality* $|\lambda_i - \lambda_i(\varepsilon)| \leq |\varepsilon|\|B\|_2$ *is valid with probability* 1 *for all* ε, *all indices* $1 \leq i \leq n$ *and for any symmetric matrix* B. *This means that* $1-\alpha$ *will almost always underestimate*

$$\mathbb{P}\left[|\lambda_i - \lambda_i(\varepsilon_2^*)| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\} \quad \textit{for all } 1 \leq i \leq n\right]$$

and therefore, from (3.21), *it will underestimate by even more the probability*

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_i(\varepsilon) \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon_2^* \textit{ and all } i \neq k].$$

Hence, ε_2^* *(given by* (3.23)*) will be rather conservative in most cases. This is confirmed theoretically, after comparison with* ε_3^* *in §3.6.*

In the next paragraph we compare ε_1^* and ε_2^* asymptotically and prove that $\varepsilon_2^* \geq \varepsilon_1^*$. This means that the second approach, that which finds ε_2^* without using the Markov's inequality, gives a better solution to Problem 3.1.1.

From Corollary 2.4.1, which was proved rigorously, when B is SGOE matrix (c.f. Definition 3.1.1) we have

$$\|B\|_2 \xrightarrow{\mathcal{D}} 2,$$

as $n \to \infty$. Therefore, when n is large, the value of $t_2^* = G_n^{-1}(1-\alpha)$ will be "close" to 2 for any α satisfying $0 < \alpha < 1$. Also, from Conjecture 2.4.3 we have that when B is defined by (3.2),

$$\mathbb{E}[\|B\|_2] \to 2,$$

as $n \to \infty$. Therefore, when n is large, we will have

$$t_2^* \approx \mathbb{E}[\|B\|_2] \approx 2.$$

On the other hand, from (3.12) and (3.23) we have that

$$\varepsilon_1^* = \frac{\alpha \min\{\gamma_{k-1}, \gamma_k\}}{2\mathbb{E}[\|B\|_2]} \qquad \text{and} \qquad \varepsilon_2^* = \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2G_n^{-1}(1-\alpha)}.$$

Therefore, using $G_n^{-1}(1-\alpha) \approx \mathbb{E}[\|B\|_2]$, we obtain

$$\frac{\varepsilon_1^*}{\varepsilon_2^*} \approx \alpha, \tag{3.24}$$

which we check in Experiment 3.9.1.

3.5. An extension to Theorem 3.2.2 and the Bauer-Fike Theorem to stochastic perturbation theory.

In this section we start with two corollaries, which stem from Theorem 3.2.2. Their aim is to provide an ε_3^*, given in Theorem 3.5.1, which solves Problem 3.1.1 and improves on ε_1^* and ε_2^*.

In particular, the following corollary gives an upper bound on the distance between a given eigenvalue in the spectrum of A and the entire spectrum of $A(\varepsilon)$. This bound is then used in Corollary 3.5.2, where we obtain an upper bound on $|\varepsilon|$ (given by (3.26)) which, as shown in Remark 3.5.1, provides a sufficient condition for $\lambda_k(\varepsilon)$ to stay separated from the rest of the eigenvalues in the spectrum of $A(\varepsilon)$. Then, in Theorem 3.5.1, all these preparatory "deterministic" results are extended to the case

of stochastic perturbations and, given a *confidence level*, $1-\alpha$, an ε_3^* is obtained, which solves Problem 3.1.1.

Firstly, we give a corollary to Theorem 3.2.2.

Corollary 3.5.1. *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices, ε be a scalar and $A(\varepsilon) = A + \varepsilon B$. Let λ_k be a simple eigenvalue of A and $\mathbf{v}_k$ be its corresponding eigenvector of unit length. Then there exists an eigenvalue, $\lambda(\varepsilon)$, from the spectrum of $A(\varepsilon)$, such that*

$$|\lambda_k - \lambda(\varepsilon)| \leq |\varepsilon| \|B\mathbf{v}_k\|_2. \tag{3.25}$$

Proof. In Theorem 3.2.2 we have to substitute $A := A(\varepsilon)$, $\mathbf{x} := \mathbf{v}_k$ and $\theta := \lambda_k$. Then, according to that theorem, we obtain

$$|\lambda_k - \lambda(\varepsilon)| \leq \|(A + \varepsilon B)\mathbf{v}_k - \lambda_k \mathbf{v}_k\|_2 = |\varepsilon| \|B\mathbf{v}_k\|_2$$

for some eigenvalue $\lambda(\varepsilon)$ from the spectrum of $A(\varepsilon)$. □

Corollary 3.5.2. *In the settings of Corollary 3.5.1, if $\hat{\varepsilon} > 0$ is such that*

$$\hat{\varepsilon}(\|B\mathbf{v}_k\| + \|B\|_2) < \min\{\gamma_{k-1}, \gamma_k\}, \tag{3.26}$$

then for all ε, such that $|\varepsilon| \leq \hat{\varepsilon}$, we have

$$|\lambda_k - \lambda_k(\varepsilon)| \leq |\varepsilon| \|B\mathbf{v}_k\| \qquad \textit{and} \qquad |\lambda_k - \lambda_j(\varepsilon)| > \hat{\varepsilon}\|B\mathbf{v}_k\| \quad \textit{for all } j \neq k.$$

Proof. Let us take $j \neq k$ and some ε, such that $|\varepsilon| \leq \hat{\varepsilon}$, and consider $|\lambda_k - \lambda_j(\varepsilon)|$. From the triangle inequality we have

$$\begin{aligned} |\lambda_k - \lambda_j(\varepsilon)| = |(\lambda_k - \lambda_j) + (\lambda_j - \lambda_j(\varepsilon))| &\geq |\lambda_k - \lambda_j| - |\lambda_j - \lambda_j(\varepsilon)| \\ &\geq \min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon| \|B\|_2 \\ &\geq \min\{\gamma_{k-1}, \gamma_k\} - \hat{\varepsilon} \|B\|_2 \end{aligned}$$

and therefore, if we assume $|\lambda_k - \lambda_j(\varepsilon)| \leq \hat{\varepsilon}\|B\mathbf{v}_k\|$, we obtain

$$\hat{\varepsilon}\|B\mathbf{v}_k\| \geq \min\{\gamma_{k-1}, \gamma_k\} - \hat{\varepsilon}\|B\|_2,$$

a contradiction with the choice of $\hat{\varepsilon}$ in (3.26).

Therefore $|\lambda_k - \lambda_j(\varepsilon)| > \hat{\varepsilon}\|B\mathbf{v}_k\|$ for all $j \neq k$. Thus, when ε satisfies $|\varepsilon| \leq \hat{\varepsilon}$ we have

$$|\lambda_k - \lambda_j(\varepsilon)| > |\varepsilon| \|B\mathbf{v}_k\| \quad \text{for all } j \neq k.$$

Hence, since (3.25) is satisfied for at least one eigenvalue from the spectrum of $A(\varepsilon)$, the only suitable candidate, which is in a distance not greater than $|\varepsilon|\|B\mathbf{v}_k\|$ from λ_k, is $\lambda_k(\varepsilon)$. □

Remark 3.5.1. *It is obvious from* (3.26) *that when we choose*

$$\hat{\varepsilon} := \frac{\min\{\gamma_{k-1}, \gamma_k\}}{\|B\mathbf{v}_k\| + \|B\|_2}, \tag{3.27}$$

then (3.26) *will be satisfied for all* ε *such that* $|\varepsilon| < \hat{\varepsilon}$. *Therefore, for all such* ε, $\lambda_k(\varepsilon)$ *will be well separated from the rest of the spectrum. In other words, for all* ε *with* $|\varepsilon| < \hat{\varepsilon}$ *and all* $j \neq k$ *we have* $\lambda_k(\varepsilon) \neq \lambda_j(\varepsilon)$.

Having done the theory in the case when A and B are deterministic matrices, we are now ready to "translate" it to the case when the matrix by which we perturb, B, is SGOE matrix (see Definition 3.1.1).

In §2.5 we prove that if B is GOE (SGOE) matrix and V is an orthogonal matrix, then V^TBV is also GOE (SGOE) matrix. Also, as we know from standard Linear Algebra, multiplication by an orthogonal matrix preserves the eigenvalues of symmetric matrices. Therefore, if

$$A = V\Lambda V^T$$

is the spectral decomposition of the symmetric matrix A, then we can consider

$$V^TA(\varepsilon)V = \Lambda + \varepsilon V^TBV, \qquad \text{instead of} \qquad A(\varepsilon) = A + \varepsilon B.$$

The eigenvalues of $V^TA(\varepsilon)V$ would be the same as those of $A(\varepsilon)$, or in probabilistic terms, the corresponding eigenvalues of $V^TA(\varepsilon)V$ and $A(\varepsilon)$ will have the same distributions. Furthermore, if B is SGOE matrix, then so would be V^TBV. Hence, without loss of generality, we may assume that

$$A(\varepsilon) = A + \varepsilon B,$$

where A is a diagonal matrix and B is SGOE matrix. In this case the eigenvectors of A, $\mathbf{v}_i$, will satisfy

$$\mathbf{v}_i = \mathbf{e}_i, \quad 1 \leq i \leq n,$$

where $\mathbf{e}_i$ is the vector whose all entries are equal to zero, apart from its i-th entry, which is equal to one. Hence, for the vector $B\mathbf{v}_k$ in Corollaries 3.5.1 and 3.5.2 we obtain

$$B\mathbf{v}_k = B\mathbf{e}_k = B(:, k),$$

where $B(:,k)$ is the MATLAB notation for the k-th column of the matrix B.

We are now ready to state the main theorem of this section.

Theorem 3.5.1. *Let A be a deterministic symmetric matrix and B be a (general) random symmetric matrix. Further, let λ_k be a simple eigenvalue of A. Finally let, given a level of confidence, $1-\alpha$, $\varepsilon_3^* > 0$ be such that*

$$\mathbb{P}\big[\varepsilon_3^*(\|B(:,k)\| + \|B\|_2) < \min\{\gamma_{k-1}, \gamma_k\}\big] = 1-\alpha. \tag{3.28}$$

Then

$$\mathbb{P}\big[|\lambda_k - \lambda_k(\varepsilon)| \leq |\varepsilon|\|B(:,k)\| \quad \textit{for all } \varepsilon,\ \textit{s.t. } |\varepsilon| \leq \varepsilon_3^*\big] \geq 1-\alpha \tag{3.29}$$

and

$$\mathbb{P}\big[|\lambda_k - \lambda_j(\varepsilon)| > \varepsilon_3^*\|B(:,k)\| \quad \textit{for all } \varepsilon,\ \textit{s.t. } |\varepsilon| \leq \varepsilon_3^* \textit{ and all } j \neq k\big] \geq 1-\alpha \tag{3.30}$$

for all ε such that $|\varepsilon| \leq \varepsilon_3^*$. Moreover,

$$\mathbb{P}[\lambda_k(\varepsilon) \neq \lambda_j(\varepsilon) \quad \textit{for all } \varepsilon,\ \textit{s.t. } |\varepsilon| \leq \varepsilon_3^* \textit{ and all } j \neq k] \geq 1-\alpha.$$

Proof. The random variable

$$X := \frac{\min\{\gamma_{k-1}, \gamma_k\}}{\|B\|_2 + \|B(:,k)\|}$$

is the solution to the linear equation

$$X(\|B\|_2 + \|B(:,k)\|) = \min\{\gamma_{k-1}, \gamma_k\}.$$

Let $F_X(t) := \mathbb{P}[X < t]$ be the c.d.f. of X. If we take

$$\varepsilon_3^* := F_X^{-1}(\alpha), \tag{3.31}$$

we would have

$$\mathbb{P}\big[\varepsilon_3^*(\|B\|_2+\|B(:,k)\|) < \min\{\gamma_{k-1}, \gamma_k\}\big] = \mathbb{P}\left[\varepsilon_3^* < \frac{\min\{\gamma_{k-1}, \gamma_k\}}{\|B\|_2 + \|B(:,k)\|}\right] = 1-F_X(\varepsilon_3^*) = 1-\alpha.$$

From Corollary 3.5.2 we have that the event

$$\{\varepsilon_3^*(\|B\|_2 + \|B(:,k)\|) < \min\{\gamma_{k-1}, \gamma_k\}\}$$

satisfies the following relations:

$$\left\{\varepsilon_3^*(\|B\|_2+\|B(:,k)\|)<\min\{\gamma_{k-1},\gamma_k\}\right\}\subset\bigcap_{\varepsilon,|\varepsilon|\leq\varepsilon_3^*}\left\{|\lambda_k-\lambda_k(\varepsilon)|\leq|\varepsilon|\|B(:,k)\|\right\} \quad (3.32)$$

and

$$\left\{\varepsilon_3^*(\|B\|_2+\|B(:,k)\|)<\min\{\gamma_{k-1},\gamma_k\}\right\}\subset\bigcap_{\varepsilon,|\varepsilon|\leq\varepsilon_3^*}\bigcap_{j\neq k}\{\lambda_k-\lambda_j(\varepsilon)|>\varepsilon_3^*\|B(:,k)\|\}. \quad (3.33)$$

Hence, by taking probabilities on both sides of (3.32) and (3.33), we obtain the inequalities in (3.29) and (3.30). Finally, since

$$\begin{aligned}\bigcap_{\varepsilon,|\varepsilon|\leq\varepsilon_3^*}\left[\bigcap_{j\neq k}\{\lambda_k-\lambda_j(\varepsilon)|>\varepsilon_3^*\|B(:,k)\|\}\quad\cap\quad\{|\lambda_k-\lambda_k(\varepsilon)|\leq|\varepsilon|\|B(:,k)\|\}\right]\\ \subset\bigcap_{\varepsilon,|\varepsilon|\leq\varepsilon_3^*}\{\lambda_k(\varepsilon)\neq\lambda_j(\varepsilon)\text{ for all }j\neq k\},\end{aligned}$$

using (3.32) and (3.33), we obtain

$$\left\{\varepsilon_3^*(\|B\|_2+\|B(:,k)\|)<\min\{\gamma_{k-1},\gamma_k\}\right\}\subset\bigcap_{\varepsilon,|\varepsilon|\leq\varepsilon_3^*}\{\lambda_k(\varepsilon)\neq\lambda_j(\varepsilon)\text{ for all }j\neq k\}$$

and thus

$$\mathbb{P}\big[\lambda_k(\varepsilon)\neq\lambda_j(\varepsilon)\quad\text{for all }\varepsilon,\text{ s.t. }|\varepsilon|\leq\varepsilon_3^*\text{ and all }j\neq k\big]\geq1-\alpha. \qquad \square$$

Remark 3.5.2. *Let us note that* (3.31) *is equivalent to*

$$\mathbb{P}\left[\frac{\min\{\gamma_{k-1},\gamma_k\}}{\|B\|_2+\|B(:,k)\|_2}<\varepsilon_3^*\right]=\alpha,$$

which is equivalent to

$$\mathbb{P}\left[\|B\|_2+\|B(:,k)\|_2>\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_3^*}\right]=\alpha$$

and thus it is also equivalent to

$$F_{\|B\|_2+\|B(:,k)\|_2}\left(\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_3^*}\right):=\mathbb{P}\left[\|B\|_2+\|B(:,k)\|_2\leq\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_3^*}\right]=1-\alpha, \quad (3.34)$$

where $F_{\|B\|_2+\|B(:,k)\|_2}(t)$ *is the c.d.f. of* $\|B\|_2 + \|B(:,k)\|_2$*. In fact,*

$$F_{\|B\|_2+\|B(:,k)\|_2}(t) := \mathbb{P}[\|B\|_2 + \|B(:,k)\|_2 < t], \tag{3.35}$$

while the inequality inside the probability in (3.34) *is not strict. However, when the distribution of* $\|B\|_2 + \|B(:,k)\|_2$ *is continuous (e.g. when* B *is SGOE matrix), the probability in* (3.34) *is equal to that in* (3.35)*, with* t *replaced by* $\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_3^*}$ *in the latter.*

Hence, (3.31) *is equivalent to*

$$\varepsilon_3^* := \frac{\min\{\gamma_{k-1},\gamma_k\}}{F^{-1}_{\|B\|_2+\|B(:,k)\|_2}(1-\alpha)}, \tag{3.36}$$

which we shall use in practice, instead of (3.31)*, because we find it easier to work with the distribution of* $\|B\|_2 + \|B(:,k)\|_2$*, than using that of* $\frac{\min\{\gamma_{k-1},\gamma_k\}}{\|B\|_2+\|B(:,k)\|_2}$*.*

We now discuss a possible way of calculating the value of ε_3^* (given in (3.36)) numerically, when B is SGOE matrix. From the proof of Theorem 3.2.2, namely from (3.31) and Remark 3.5.2, we can see that ε_3^* can be obtained by inverting the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$. Using the results in §2 we can calculate numerically the c.d.f. of $\|B\|_2$. Also, from the definition of SGOE matrix (c.f. Definition 3.1.1) we know that

$$\|B(:,k)\|_2^2 = \sum_{i\neq k} B_{ik}^2 + B_{kk}^2 =: \xi + \eta,$$

where ξ and η are independent random variables satisfying $n\xi \in \chi^2_{n-1}$ and $\frac{n}{2}\eta \in \chi^2_1$. (Here by χ^2_i we have denoted the Chi-square distribution with i degrees of freedom.) But the difficulty in calculating the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$ comes from the fact that we don't know how these two random variables depend on each other. In other words, we don't know their joint distribution. However, as we discuss in the next paragraph, we can assume that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent. The implications of such an assumption are that we shall be able to find the c.d.f. $F_{\|B\|_2+\|B(:,k)\|_2}$ numerically and so we shall also be able to invert it. Thus, given a confidence level, $1-\alpha$, we shall be able to find ε_3^* from (3.36) (or equivalently, from (3.31)).

Conjecture 3.5.1. *For the sake of ease of numerical computation, we* ***assume*** *that* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent random variables.*

The question is whether the assumption that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent is at least "approximately valid". Intuitively, the value of $\|B\|_2$ depends on the elements of B above and on its main diagonal. Hence, $\|B\|_2$ depends on $\frac{n^2+n}{2}$ parameters. On the

other hand, the value of $\|B(:,k)\|_2$ depends only on n of these parameters. Therefore, as n becomes larger, the value of $\|B\|_2$ should become "less dependent" on the value of $\|B(:,k)\|_2$. Thus, this makes our assumption "reasonable", when n is sufficiently large. Numerical tests (see Experiment 3.9.2) are consistent with this assumption.

3.6. Asymptotic comparison between ε_2^* and ε_3^*.

The main result in this section is to show that (asymptotically) $\varepsilon_3^* > \varepsilon_2^*$ and thus ε_3^* provides a better solution to Problem 3.1.1. We prove this using standard results in Probability Theory, including Slutsky's Theorem (c.f. Theorem A.1.7).

Firstly, we have the following lemma.

Lemma 3.6.1. *Let $X_n = (x_1, x_2, \ldots, x_n)$, where x_i are i.i.d. random variables distributed $\mathcal{N}(0,1)$. Then*

$$\frac{1}{n}\|X_n\|_2^2 \xrightarrow{\mathcal{D}} 1, \quad \text{as } n \to \infty.$$

Proof. We apply the Law of large numbers (c.f. Theorem A.1.9) to the sequence $\{x_i^2\}_{i\in\mathbb{N}}$ of i.i.d. random variables. We have $\mathbb{E}[x_i^2] = 1 < \infty$ and

$$\mathrm{Var}[x_i^2] = \mathbb{E}[x_i^4] - \big(\mathbb{E}[x_i^2]\big)^2 = 3 - 1 = 2$$

(see Example A.1.2 for the calculation of $\mathbb{E}[x_i^4]$). Therefore, from the Law of large numbers we infer that

$$\frac{1}{n}\|X_n\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{\mathcal{D}} 1$$

(see Remark A.1.6 after Theorem A.1.9). □

Lemma 3.6.2. *Let $B \in \mathbb{R}^{n\times n}$ be SGOE matrix. Then if the vector $\mathbf{b}$ represents a row or a column of B, we have*

$$\|\mathbf{b}\| \xrightarrow{\mathcal{D}} 1 \qquad \textit{and (thus)} \qquad \|\mathbf{b}\| \xrightarrow{\mathbb{P}} 1,$$

as $n \to \infty$.

Remark 3.6.1. *For the relation between* convergence in distribution*, written $\xrightarrow{\mathcal{D}}$, and* convergence in probability*, written $\xrightarrow{\mathbb{P}}$, see Definition A.1.16 and Theorem A.1.5.*

Proof. Without loss of generality we may assume that $\mathbf{b}$ is the first column of the matrix B. From the definition of the matrix B (see (3.2)) we know that the components of the

vector $\mathbf{b}$, $\mathbf{b}_i$ with $1 \leq i \leq n$, are independent, normally distributed random variables. More precisely, for the entries $\mathbf{b}_i$, $1 \leq i \leq n$, of $\mathbf{b}$ we have

$$\mathbf{b}_1 \in \mathcal{N}\left(0, \frac{2}{n}\right) \qquad \text{and} \qquad \mathbf{b}_i \in \mathcal{N}\left(0, \frac{1}{n}\right) \quad \text{for } 2 \leq i \leq n.$$

Let us define the vector $\bar{\mathbf{b}}$ in the following way:

$$\bar{\mathbf{b}}_1 := \frac{1}{\sqrt{2}}\mathbf{b}_1 \qquad \text{and} \qquad \bar{\mathbf{b}}_i := \mathbf{b}_i \quad \text{for } 2 \leq i \leq n.$$

Then the vector $\bar{\mathbf{b}}$ is the same as the vector $\frac{1}{\sqrt{n}}X_n$ in Lemma 3.6.1 and thus

$$\|\bar{\mathbf{b}}\|^2 \xrightarrow{\mathcal{D}} 1,$$

as $n \to \infty$. The latter implies

$$\|\bar{\mathbf{b}}\| \xrightarrow{\mathcal{D}} 1 \qquad \text{and therefore} \qquad \|\bar{\mathbf{b}}\| \xrightarrow{\mathbb{P}} 1,$$

as $n \to \infty$. Further, using the chain of inequalities

$$\begin{aligned} |\|\mathbf{b}\| - 1| \leq \left|\|\mathbf{b}\| - \|\bar{\mathbf{b}}\|\right| + \left|\|\bar{\mathbf{b}}\| - 1\right| &\leq \|\mathbf{b} - \bar{\mathbf{b}}\| + \left|\|\bar{\mathbf{b}}\| - 1\right| \\ &= \frac{\sqrt{2}-1}{\sqrt{2}}|\mathbf{b}_1| + \left|\|\bar{\mathbf{b}}\| - 1\right| \end{aligned}$$

and the fact that for any $\delta > 0$

$$\mathbb{P}\left[\frac{\sqrt{2}-1}{\sqrt{2}}|\mathbf{b}_1| > \delta\right] = \mathbb{P}\left[|\mathbf{b}_1| > \frac{\delta\sqrt{2}}{\sqrt{2}-1}\right] = \frac{1}{\sqrt{2\pi}}\int_{\frac{\delta\sqrt{n}}{\sqrt{2}-1}}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt \to 0,$$

as $n \to \infty$, we can infer that for any $\delta > 0$

$$\begin{aligned} \mathbb{P}\left[|\|\mathbf{b}\| - 1| > \delta\right] &\leq \mathbb{P}\left[\frac{\sqrt{2}-1}{\sqrt{2}}|\mathbf{b}_1| + \left|\|\bar{\mathbf{b}}\| - 1\right| > \delta\right] \\ &\leq \mathbb{P}\left[\frac{\sqrt{2}-1}{\sqrt{2}}|\mathbf{b}_1| > \frac{\delta}{2}\right] + \mathbb{P}\left[\left|\|\bar{\mathbf{b}}\| - 1\right| > \frac{\delta}{2}\right] \to 0, \end{aligned}$$

as $n \to \infty$. In other words $\|\mathbf{b}\| \xrightarrow{\mathbb{P}} 1$, as $n \to \infty$. Hence $\|\mathbf{b}\| \xrightarrow{\mathcal{D}} 1$, as $n \to \infty$. □

We now combine the results of Lemmas 3.6.1 and 3.6.2, Slutsky's theorem (c.f. Theorem A.1.7) and Corollary 2.4.1 to arrive at the following results.

Proposition 3.6.1. *Let B be SGOE (see Definition 3.1.1) and let the number* $\min\{\gamma_{k-1},\gamma_k\}$ *be independent of n. Then we obtain*

$$\frac{\min\{\gamma_{k-1},\gamma_k\}}{2\|B\|_2} \xrightarrow{\mathcal{D}} \frac{\min\{\gamma_{k-1},\gamma_k\}}{4}, \tag{3.37}$$

The implication of Proposition 3.6.1 is that, since ε_2^* is found from

$$1-\alpha = \mathbb{P}\left[\varepsilon_2^*\|B\|_2 < \frac{1}{2}\min\{\gamma_{k-1},\gamma_k\}\right] = \mathbb{P}\left[\frac{\min\{\gamma_{k-1},\gamma_k\}}{2\|B\|_2} > \varepsilon_2^*\right],$$

(see (3.22)) then $\varepsilon_2^* \approx \frac{1}{4}\min\{\gamma_{k-1},\gamma_k\}$ when n is sufficiently large.

Proposition 3.6.2. *Let B be SGOE (see Definition 3.1.1) and let the number* $\min\{\gamma_{k-1},\gamma_k\}$ *be independent of n. Then we obtain*

$$\frac{\min\{\gamma_{k-1},\gamma_k\}}{\|B(:,k)\|_2+\|B\|_2} \xrightarrow{\mathcal{D}} \frac{\min\{\gamma_{k-1},\gamma_k\}}{3}, \tag{3.38}$$

as $n\to\infty$.

From Proposition 3.6.2 we can conclude that, since ε_3^* is calculated from

$$\mathbb{P}\left[\frac{\min\{\gamma_{k-1},\gamma_k\}}{\|B\|_2+\|B(:,k)\|_2} > \varepsilon_3^*\right] = 1-\alpha,$$

(see (3.28)) then $\varepsilon_3^* \approx \frac{1}{3}\min\{\gamma_{k-1},\gamma_k\}$ when n is sufficiently large. Hence,

$$\frac{\varepsilon_2^*}{\varepsilon_3^*} \approx \frac{3}{4}. \tag{3.39}$$

This means that ε_3^* is a better solution to Problem 3.1.1, but it has the disadvantage that its calculation involves the inversion of the c.d.f. of $\|B\|_2+\|B(:,k)\|_2$ (see the discussion at the end of §3.5 for further details).

3.7. Stochastic version of the $\sin\psi$ Theorem.

Let us recall that in this chapter we use the following notation: $A(\varepsilon) := A+\varepsilon B$, where $A, B \in \mathbb{R}^{n\times n}$ are symmetric matrices and $\varepsilon\in\mathbb{R}$; Also $\lambda_1(\varepsilon)\le\lambda_2(\varepsilon)\le\cdots\le\lambda_n(\varepsilon)$ are the eigenvalues of $A(\varepsilon)$, $\mathbf{v}_1(\varepsilon),\mathbf{v}_2(\varepsilon),\ldots,\mathbf{v}_n(\varepsilon)$ are their corresponding unit eigenvectors and $\lambda_i=\lambda_i(0)$ and $\mathbf{v}_i=\mathbf{v}_i(0)$ for $1\le i\le n$. The eigenvalue λ_k of A is assumed simple in this section. Finally, $\gamma_i(\varepsilon) := \lambda_{i+1}(\varepsilon)-\lambda_i(\varepsilon)$ for $i=1,2,\ldots,n-1$ with $\gamma_i=\gamma_i(0)$.

Since we are interested in the angle between $\mathbf{v}_k$ and $\mathbf{v}_k(\varepsilon)$, we shall restate Theorem 3.2.4 so that $\angle(\mathbf{y},\mathbf{v}_k)$ is replaced by $\angle(\mathbf{v}_k,\mathbf{v}_k(\varepsilon))$. Firstly, we replace A, λ_k and $\mathbf{v}_k$

in Theorem 3.2.4 by $A(\varepsilon)$, $\lambda_k(\varepsilon)$ and $\mathbf{v}_k(\varepsilon)$ respectively, where $\lambda_k(\varepsilon)$ is the eigenvalue in the spectrum of $A(\varepsilon)$, which stays in a closest distance to the number θ. Secondly, we shall determine the number θ and the unit vector $\mathbf{y}$, which were arbitrary in Theorem 3.2.4. In Corollary 3.7.1 we choose $\theta := \lambda_k$ and $\mathbf{y} := \mathbf{v}_k$. However, care should be taken for θ to be closer to $\lambda_k(\varepsilon)$ than to any other eigenvalue in the spectrum of $A(\varepsilon)$. This condition is fulfilled with the help of Bauer-Fike's Theorem (c.f. Corollary 3.2.1). It states that $|\lambda_k - \lambda_k(\varepsilon)| \leq \|\varepsilon B\|_2$. Therefore, when $\theta := \lambda_k$ we require $\|\varepsilon B\|_2 \leq \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}$.

Corollary 3.7.1. *Let λ_k be a simple eigenvalue of the matrix A and $\mathbf{v}_k$ be its corresponding unit eigenvector. Also let*

$$|\varepsilon|\left(\|B\|_2 + \|B(:,k)\|_2\right) < \min\{\gamma_{k-1}, \gamma_k\}.$$

Then $\lambda_k(\varepsilon)$ is the eigenvalue in the spectrum of $A(\varepsilon)$, which is closest to λ_k. Moreover, if $\psi := \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$, the following inequality holds

$$|\sin\psi| \leq \frac{|\varepsilon|\|B\mathbf{v}_k\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2}. \tag{3.40}$$

Proof. In Corollary 3.5.2 we proved that when $|\varepsilon|(\|B\|_2 + \|B(:,k)\|_2) < \min\{\gamma_{k-1}, \gamma_k\}$, then

$$|\lambda_k - \lambda_k(\varepsilon)| \leq |\varepsilon|\|B\mathbf{v}_k\|_2 \qquad \text{and} \qquad |\lambda_k - \lambda_j(\varepsilon)| > |\varepsilon|\|B\mathbf{v}_k\|_2, \quad \text{for all } j \neq k. \tag{3.41}$$

We now apply Theorem 3.2.4 with A, λ_k and $\mathbf{v}_k$ replaced by $A(\varepsilon)$, $\lambda_k(\varepsilon)$ and $\mathbf{v}_k(\varepsilon)$, respectively. The constant θ and the vector $\mathbf{y}$, which were arbitrary in Theorem 3.2.4, are now defined as $\theta := \lambda_k$ and $\mathbf{y} := \mathbf{v}_k$. Then, it follows from from (3.41) that

$$\min_{1\leq i\leq n} |\lambda_i(\varepsilon) - \lambda_k| = |\lambda_k(\varepsilon) - \lambda_k|.$$

Therefore, the angle $\psi = \angle(\mathbf{y}, \mathbf{v}_k)$ in Theorem 3.2.4 now becomes $\psi = \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$ and the residual, $r(\mathbf{y}) = A\mathbf{y} - \theta\mathbf{y}$ there, now becomes

$$r(\mathbf{v}_k) = (A + \varepsilon B)\mathbf{v}_k - \lambda_k\mathbf{v}_k = \varepsilon B\mathbf{v}_k.$$

Therefore, the result of Theorem 3.2.4 in the new settings becomes

$$|\sin\psi| \leq \frac{\|\varepsilon B\mathbf{v}_k\|_2}{\gamma_\theta}.$$

In order to bound the quantity γ_θ from below, we use the second part of (3.41), which implies

$$\gamma_\theta = |\lambda_k - \lambda_{j_0}(\varepsilon)| \qquad \text{for some } j_0 \neq k.$$

Therefore

$$\gamma_\theta \geq |\lambda_k - \lambda_{j_0}| - |\lambda_{j_0} - \lambda_{j_0}(\varepsilon)| \geq \min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2,$$

where in order to get the last inequality we have used $|\lambda_k - \lambda_{j_0}| \geq \min\{\gamma_{k-1}, \gamma_k\}$ and the Bauer-Fike Theorem, $|\lambda_{j_0} - \lambda_{j_0}(\varepsilon)| \leq |\varepsilon|\|B\|_2$.

Hence, we finally obtain

$$|\sin\psi| \leq \frac{|\varepsilon|\|B\mathbf{v}_k\|}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|},$$

as required. □

Remark 3.7.1. *When ε is such that $|\varepsilon|(\|B\|_2 + \|B(:,k)\|_2) < \min\{\gamma_{k-1}, \gamma_k\}$, the upper bound on $|\sin\psi|$ in* (3.40)*,*

$$|\sin\psi| \leq \frac{|\varepsilon|\|B\mathbf{v}_k\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2},$$

implies $|\sin\psi| < 1$, as one should expect.

Now we state Theorem 3.2.5, which provides a lower bound on $|\sin\psi|$, in such a way that the angle $\psi = \angle(\mathbf{y}, \mathbf{z})$ in that theorem becomes $\psi = \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$. This is done in the next corollary.

Corollary 3.7.2. *In the settings of this chapter,*

$$|\sin\psi| \geq \frac{\|(I - \mathbf{v}_k\mathbf{v}_k^T)B\mathbf{v}_k\|_2}{\text{spread}(A) + 2|\varepsilon|\|B\|_2}, \tag{3.42}$$

where $\psi := \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$.

Proof. If we replace A, α and $\mathbf{z}$ in Theorem 3.2.5 with $A(\varepsilon)$, $\lambda_k(\varepsilon)$ and $\mathbf{v}_k(\varepsilon)$, respectively and let $\mathbf{y} := \mathbf{v}_k$ we obtain

$$\theta = \mathbf{v}_k^T(A + \varepsilon B)\mathbf{v}_k = \lambda_k + \varepsilon\mathbf{v}_k^T B\mathbf{v}_k,$$

which leads to

$$r(\mathbf{v}_k) = (A + \varepsilon B)\mathbf{v}_k - (\lambda_k + \varepsilon\mathbf{v}_k^T B\mathbf{v}_k)\mathbf{v}_k = (I - \mathbf{v}_k\mathbf{v}_k^T)B\mathbf{v}_k$$

and hence (3.8) becomes

$$|\sin\psi| \geq \frac{\|(I - \mathbf{v}_k\mathbf{v}_k^T)B\mathbf{v}_k\|_2}{\text{spread}(A(\varepsilon))}.$$

Here we use

$$\text{spread}(A(\varepsilon)) = \lambda_n(\varepsilon) - \lambda_1(\varepsilon) \leq \lambda_n + |\varepsilon|\|B\|_2 - \lambda_1 + |\varepsilon|\|B\|_2 = \text{spread}(A) + 2|\varepsilon|\|B\|_2,$$

which finally gives us (3.42). □

The results so far hold for any pair of symmetric matrices, A and B. If $A = V\Lambda V^T$ is the spectral decomposition of A and

$$\tilde{A}(\varepsilon) := \Lambda + \varepsilon V^T B V, \tag{3.43}$$

we have $A(\varepsilon) = V\tilde{A}(\varepsilon)V^T$. Let $\tilde{\lambda}_1(\varepsilon) \leq \tilde{\lambda}_2(\varepsilon) \leq \cdots \leq \tilde{\lambda}_n(\varepsilon)$ be the eigenvalues of $\tilde{A}(\varepsilon)$ and $\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_n$ be their corresponding unit eigenvectors. Therefore we have $\tilde{\lambda}_i(\varepsilon) = \lambda_i(\varepsilon)$ and $\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon)) = \angle(\mathbf{e}_k, \tilde{\mathbf{v}}_k(\varepsilon))$, where $\mathbf{e}_k$ is the vector whose entries are zeros, apart from its k-th entry, which is equal to one.

Let B be SGOE matrix. Then the eigenvalues of $A(\varepsilon)$, $\lambda_i(\varepsilon)$, $1 \leq i \leq n$, are random variables and so is also $|\sin\psi|$, where $\psi = \angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$. Since $\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon)) = \angle(\mathbf{e}_k, \tilde{\mathbf{v}}_k(\varepsilon))$ in the case of deterministic perturbations, then, when the matrix B is random, the distributions of $|\sin\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))|$ and $|\sin\angle(\mathbf{e}_k, \tilde{\mathbf{v}}_k(\varepsilon))|$ should be identical. Therefore, from (3.43), the distribution of $|\sin\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))|$ depends only on $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the distribution of the entries of the matrix $V^T B V$. But in Corollary 2.5.1 we showed that if B is SGOE matrix, then $V^T B V$ is also SGOE matrix. Hence, the distribution of $|\sin\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))|$, and in fact the distribution of any measurable function of $\angle(\mathbf{v}_k, \mathbf{v}_k(\varepsilon))$, doesn't depend on the choice of the matrix V, whose columns are the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Therefore, without loss of generality, we may consider only perturbations of the form $A + \varepsilon B$, where A is a diagonal matrix and B is SGOE matrix.

Before we state our next results, which are extensions to Corollaries 3.7.1 and 3.7.2 to the case of stochastic perturbation, we shall rewrite the bounds on $|\sin\psi|$, assuming the matrix A is diagonal. In this new settings the eigenvectors of A, $\mathbf{v}_i$, satisfy $\mathbf{v}_i = \mathbf{e}_i$, $1 \leq i \leq n$. Therefore

$$B\mathbf{v}_k = B\mathbf{e}_k = B(:,k) \qquad \text{and} \qquad (I - \mathbf{v}_k\mathbf{v}_k^T)B\mathbf{v}_k = B(:,k) - B_{kk}\mathbf{e}_k,$$

where we recall that the notation $B(:,k)$ refers to the k-th column of the matrix B.

We now restate Corollaries 3.7.1 and 3.7.2 in the case when B is SGOE matrix.

Corollary 3.7.3. *Let λ_k be a simple eigenvalue of the diagonal matrix A and B be SGOE matrix. Let also $\mathbf{v}_k(\varepsilon)$ be the unit eigenvector of $A(\varepsilon)$ corresponding to $\lambda_k(\varepsilon)$ and let the angle $\psi := \angle(\mathbf{e}_k, \mathbf{v}_k(\varepsilon))$. Finally, let ε be such that*

$$\mathbb{P}\big[|\varepsilon|(\|B\|_2 + \|B(:,k)\|_2) < \min\{\gamma_{k-1}, \gamma_k\}\big] = 1 - \alpha$$

for some α satisfying $0 \leq \alpha \leq 1$. Then

$$\mathbb{P}\left[|\sin\psi| \leq \frac{|\varepsilon|\|B(:,k)\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2}\right] \geq 1 - \alpha. \tag{3.44}$$

Proof. When we interpret Corollary 3.7.1 in probabilistic terms, we have the following relation in terms of events

$$\big\{|\varepsilon|(\|B\|_2 + \|B(:,k)\|_2) < \min\{\gamma_{k-1}, \gamma_k\}\big\} \subset \left\{|\sin\psi| \leq \frac{|\varepsilon|\|B(:,k)\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2}\right\}.$$

Therefore, after taking probabilities on both sides, the proof is complete. □

Corollary 3.7.4. *Let A be a diagonal matrix and B be SGOE matrix. Let also $\psi := \angle(\mathbf{e}_k, \mathbf{v}_k(\varepsilon))$. Then the following inequality holds with probability one*

$$|\sin\psi| \geq \frac{\|B(:,k) - B_{kk}\mathbf{e}_k\|_2}{\mathrm{spread}(A) + 2|\varepsilon|\|B\|_2}. \tag{3.45}$$

We now find the limits, as $n \to \infty$, of the bounds of $|\sin\psi|$, given in Corollaries 3.7.1 and 3.7.2. In order to do that we use the results in Corollaries 3.7.3 and 3.7.4, combined with Lemma 3.6.2 and Slutsky's theorem (c.f. Theorem A.1.7). We also use $\|B\|_2 \xrightarrow{\mathcal{D}} 2$, as $n \to \infty$, when B is defined by (3.2). The results are presented in Corollary 3.7.5. They may be useful when one wants to approximate the bounds on the distribution of $|\sin\psi|$ when n is large and can not find the exact distributions of the bounds in Corollaries 3.7.3 and 3.7.4.

Corollary 3.7.5. *Let B be SGOE matrix and the scalars ε and $\min\{\gamma_{k-1}, \gamma_k\}$ are independent of n. Further, let $|\varepsilon| < \frac{1}{2}\min\{\gamma_{k-1}, \gamma_k\}$. Then the following convergence results hold:*

(a) $\dfrac{|\varepsilon|\|B(:,k)\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - |\varepsilon|\|B\|_2} \xrightarrow{\mathcal{D}} \dfrac{|\varepsilon|}{\min\{\gamma_{k-1}, \gamma_k\} - 2|\varepsilon|}$, *as $n \to \infty$;*

(b) $\dfrac{|\varepsilon|\|B(:,k) - B_{kk}\mathbf{e}_k\|_2}{\mathrm{spread}(A) + 2|\varepsilon|\|B\|_2} \xrightarrow{\mathcal{D}} \dfrac{|\varepsilon|}{\mathrm{spread}(A) + 4|\varepsilon|}$, *as $n \to \infty$.*

Proof. We shall only prove (b), as the proof of (a) is similar. Using Lemma 3.6.1, we obtain

$$\|B(:,k) - B_{kk}\mathbf{e}_k\|_2 \xrightarrow{\mathcal{D}} 1, \quad \text{as } n \to \infty,$$

since

$$n\|B(:,k) - B_{kk}\mathbf{e}_k\|_2^2 \in \chi^2_{n-1}.$$

Also, we know from §2.3 that $\|B\|_2 \xrightarrow{\mathcal{D}} 2$, as $n \to \infty$, when B is SGOE matrix. Therefore, we can apply Slutsky's theorem (see Theorem A.1.7) to the sequence

$$\left\{ \frac{|\varepsilon|\|B(:,k) - B_{kk}\mathbf{e}_k\|_2}{\text{spread}(A) + 2|\varepsilon|\|B\|_2} \right\}_{n\in\mathbb{N}}$$

and conclude

$$\frac{|\varepsilon|\|B(:,k) - B_{kk}\mathbf{e}_k\|_2}{\text{spread}(A) + 2|\varepsilon|\|B\|_2} \xrightarrow{\mathcal{D}} \frac{|\varepsilon|}{\text{spread}(A) + 4|\varepsilon|},$$

as $n \to \infty$. □

Remark 3.7.2. *All results in Corollary 3.7.5 may be stated as convergence in probability. In general convergence in probability implies convergence in distribution, without the inverse necessarily being true. However, when a sequence of random variables converges in distribution to a constant, the two types of convergence are equivalent.*

The research on the sensitivity of *individual components* of eigenvectors, perturbed by deterministic symmetric (or SGOE) matrices, is still in its preliminary stage, and the results we have proved here are the only tools we know of. Next we discuss a possible application of these results.

We showed that, if $\psi := \angle(\mathbf{v}, \mathbf{v}(\varepsilon))$, then $|\sin\psi|$ is bounded in the following way:

$$|\sin\psi| \leq \frac{|\varepsilon|\|B(:,k)\|_2}{\min\{\gamma_{k-1}, \gamma_k\} - \varepsilon\|B\|_2}. \tag{3.46}$$

Let $\mathbf{v}_k^{[j]}$ and $\mathbf{v}_k^{[j]}(\varepsilon)$ be the j-th component of $\mathbf{v}_k$ and $\mathbf{v}_k(\varepsilon)$, respectively, and

$$\boldsymbol{\delta}(\varepsilon) := \mathbf{v}(\varepsilon) - \mathbf{v}.$$

We measure the sensitivity of the j-th component of $\mathbf{v}_k$, $\mathbf{v}_k^{[j]}$, to perturbation, by considering the difference

$$\delta_j(\varepsilon) := \mathbf{v}_k^{[j]}(\varepsilon) - \mathbf{v}_k^{[j]} = \mathbf{e}_j^T(\mathbf{v}_k(\varepsilon) - \mathbf{v}_k).$$

From here, using the cosine rule, it is straightforward to show that

$$\|\boldsymbol{\delta}(\varepsilon)\|^2 = \|\mathbf{v}_k\|^2 + \|\mathbf{v}_k(\varepsilon)\|^2 - 2\|\mathbf{v}_k\|\|\mathbf{v}_k(\varepsilon)\|\cos\psi = 2(1-\cos\psi) = 4\sin^2\frac{\psi}{2}.$$

Then, noting that $\sin^2 x = \sin^2|x|$ and using the inequality $\sin x \leq x$ for all $x \geq 0$, we obtain

$$\|\boldsymbol{\delta}\| \leq |\psi|$$

and therefore, from (3.46) and from the fact that arcsin is an increasing function, we further get

$$\|\boldsymbol{\delta}(\varepsilon)\| \leq |\psi| \leq \arcsin\frac{|\varepsilon|\|B(:,k)\|}{\min\{\gamma_{k-1},\gamma_k\} - |\varepsilon|\|B\|},$$

and thus

$$|\delta_j(\varepsilon)| \leq \arcsin\frac{|\varepsilon|\|B(:,k)\|}{\min\{\gamma_{k-1},\gamma_k\} - |\varepsilon|\|B\|} \tag{3.47}$$

For the second inequality we have used that $|\delta_j(\varepsilon)| = |\mathbf{e}_j^T\boldsymbol{\delta}(\varepsilon)| \leq \|\boldsymbol{\delta}(\varepsilon)\|$ for any index $1 \leq j \leq n$. The equality $|\mathbf{e}_j^T\boldsymbol{\delta}(\varepsilon)| = \|\delta(\varepsilon)\|$ takes place if and only if, $\mathbf{e}_j$ and $\boldsymbol{\delta}(\varepsilon)$ are collinear. This collinearity between $\mathbf{e}_j$ and $\boldsymbol{\delta}(\varepsilon)$, in terms of the eigenvector $\mathbf{v}_k$, means that only its j-th entry is perturbed. In other words, the considerations in this paragraph may be useful for bounding from above the perturbation to the most sensitive element of $\mathbf{v}_k$.

Further, by the Mean-value Theorem we have

$$\mathbf{v}_k(\varepsilon) - \mathbf{v}_k = \varepsilon\mathbf{v}_k'(x_\varepsilon),$$

where $x_\varepsilon \in (0,\varepsilon)$ and ε is sufficiently small. Hence

$$|\delta_j(\varepsilon)| = |\varepsilon||\mathbf{e}_j^T\mathbf{v}_k'(x_\varepsilon)| \quad \text{for some } x_\varepsilon \in (0,\varepsilon).$$

Thus, loosely speaking, if $\mathbf{v}_k^{[j]}$ is the most sensitive element of $\mathbf{v}_k$, that is, the element for which $\left|\mathbf{v}_k^{[i]\prime}(x_\varepsilon)\right| \leq \left|\mathbf{v}_k^{[j]\prime}(x_\varepsilon)\right|$ for all $i \neq j$, then $\mathbf{e}_j$ and $\mathbf{v}_k'(x_\varepsilon)$ are "almost" collinear.

The discussion above is a first step towards linking the results for $|\sin\psi|$, presented in this section, with the analysis of sensitivity to perturbation of the elements of a given eigenvector. For example, by (3.47) we can bound from above the perturbations in the most sensitive element of some eigenvector. In our future work, we shall aim to refine the bound in (3.47). Also, we shall seek to extend the results of this section to bounds on the perturbation of eigenspaces.

3.8. An extension to Theorem 3.2.2 and the Bauer-Fike Theorem to stochastic perturbations of rectangular matrices.

In this section we briefly mention a way, in which the theory in §3.5 could be extended to the case of stochastic perturbations to rectangular matrices. As such these perturbation results will have application to perturbation of spectral clustering of micro-array data, using singular vectors, as discussed by (Higham et al., 2007). We are interested in perturbations of the form

$$A(\varepsilon) := A + \varepsilon W, \tag{3.48}$$

where $A, A(\varepsilon)$ and W are rectangular $n \times p$ matrices, where $n > p$. More precisely, as in §3.5, given a *confidence level*, $1 - \alpha$, we shall find an ε_r^* such that, if σ_k is a simple singular value of A, then $\sigma_k(\varepsilon)$ is also a simple eigenvalue of $A(\varepsilon)$, for all ε satisfying $|\varepsilon| \leq \varepsilon_r^*$.

The matrix by which we perturb A, W, will be assumed deterministic until Theorem 3.8.1, where it is assumed a general random matrix. At the end of this section we discuss a possible application of Theorem 3.8.1 for a specific class of random matrices, to which W belongs.

In this section the singular values of A will be denoted by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ and their corresponding left and right singular vectors, by $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$, respectively. Further, all singular vectors of A are assumed of unit length. We start with a key lemma, which shall link perturbations to singular values of rectangular matrices to perturbations to the symmetric eigenvalue problem.

Lemma 3.8.1. *Let A be any $n \times p$ rectangular matrix with singular values and singular vectors as defined above. Let us define the $(n+p) \times (n+p)$ symmetric matrix*

$$\check{A} := \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \tag{3.49}$$

Then the spectra of $\check{A}$ are as follows:

(a) $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ are eigenvalues of $\check{A}$, with corresponding unit eigenvectors $\frac{1}{\sqrt{2}}[\mathbf{u}_i, \mathbf{v}_i]^T$, $1 \leq i \leq p$;

(b) $-\sigma_1 \leq -\sigma_2 \leq \cdots \leq -\sigma_p$ are eigenvalues of $\check{A}$, with corresponding unit eigenvectors $\frac{1}{\sqrt{2}}[-\mathbf{u}_i, \mathbf{v}_i]^T$, $1 \leq i \leq p$;

(c) 0 is an eigenvalue of $\check{A}$ with corresponding eigenvectors $[\mathbf{u}_{p+1}, 0]^T, [\mathbf{u}_{p+2}, 0]^T, \ldots [\mathbf{u}_n, 0]^T$.

Proof. The proof can be done by checking directly that these are indeed eigenvalues and eigenvectors of $\check{A}$. Further, one can easily show that the eigenvectors of $\check{A}$ form an orthonormal basis in $\mathbb{R}^{n+p}$. □

The following result is an easy consequence of Lemma 3.8.1. It shall be needed later in this section, since it expresses the 2-norm of $\check{A}$ via the spectrum of A.

Corollary 3.8.1. *Let the symmetric matrix $\check{A}$ be defined by* (3.49)*, where A is some $n \times p$ rectangular matrix with singular values and singular vectors as defined above. Then $\|\check{A}\|_2 = \sigma_1$ and therefore $\|\check{A}\|_2 = \|A\|_2$.*

Proof. From Lemma 3.8.1 we can see that σ_1 is the eigenvalue of $\check{A}$ with the largest magnitude. Therefore, by the definition of the 2-norm, we have $\|\check{A}\|_2 = \sigma_1$. The second claim of the corollary, that $\|\check{A}\|_2 = \|A\|_2$, follows from what we already proved, that $\|\check{A}\|_2 = \sigma_1$, and the definition of the 2-norm for rectangular matrices. □

Before we proceed, let us make some final preparations. For the rest of this section we shall assume that the singular value σ_k is positive (nonzero) and simple, where k is some index, $1 \leq k \leq p$. We shall also denote

$$\gamma_i := \sigma_i - \sigma_{i+1}, \quad 1 \leq i \leq p-1 \qquad \text{and} \qquad \gamma_p := \sigma_p.$$

Given a matrix $W \in \mathbb{R}^{n\times p}$ and a scalar ε, the matrix $A(\varepsilon)$ shall be defined by (3.48) and its singular values will be denoted by $\sigma_1(\varepsilon) \geq \sigma_2(\varepsilon) \geq \cdots \geq \sigma_p(\varepsilon)$. Finally, in these settings we define

$$\mathbf{r}_i := \frac{1}{\sqrt{2}}[W\mathbf{v}_i, W^T\mathbf{u}_i]^T, \quad 1 \leq i \leq p.$$

We now restate Corollary 3.5.2 and Theorem 3.5.1 in terms of rectangular matrices. In doing this we shall use Lemma 3.8.1.

Proposition 3.8.1. *Let $A, W \in \mathbb{R}^{n\times p}$ be rectangular matrices and $A(\varepsilon)$ be defined by* (3.48)*. If $\hat{\varepsilon} > 0$ is such that*

$$\hat{\varepsilon}\,(\|W\|_2 + \|\mathbf{r}_k\|_2) < \min\{\gamma_{k-1}, \gamma_k\},$$

then for all ε, such that $|\varepsilon| \leq \hat{\varepsilon}$, we have

$$|\sigma_k - \sigma_k(\varepsilon)| \leq |\varepsilon|\|\mathbf{r}_k\|_2 \qquad \text{and} \qquad |\sigma_k - \sigma_i(\varepsilon)| > \hat{\varepsilon}\|\mathbf{r}_k\|_2 \quad \text{for all } i \neq k.$$

Proof. Similarly to $\check{A}$, let us define the symmetric $(n+p)\times(n+p)$ matrices, $\check{W}$ and

$\check{A}(\varepsilon)$, by

$$\check{W} := \begin{bmatrix} 0 & W \\ W^T & 0 \end{bmatrix} \qquad \text{and} \qquad \check{A}(\varepsilon) := \begin{bmatrix} 0 & A(\varepsilon) \\ A(\varepsilon)^T & 0 \end{bmatrix}. \tag{3.50}$$

Then $\check{A}(\varepsilon) = \check{A} + \varepsilon\check{W}$ and the eigenvalues of $\check{A}$ and $\check{A}(\varepsilon)$ are

$$-\sigma_1 \leq -\sigma_2 \leq \cdots \leq -\sigma_p \leq 0 \leq \sigma_p \leq \cdots \leq \sigma_2 \leq \sigma_1$$

and

$$-\sigma_1(\varepsilon) \leq -\sigma_2(\varepsilon) \leq \cdots \leq -\sigma_p(\varepsilon) \leq 0 \leq \sigma_p(\varepsilon) \leq \cdots \leq \sigma_2(\varepsilon) \leq \sigma_1(\varepsilon),$$

respectively, with σ_k being a simple eigenvalue of $\check{A}$. Therefore we can apply Corollary 3.5.2 to $\check{A}$, $\check{W}$, $\check{A}(\varepsilon)$ and σ_k, instead of A, B, $A(\varepsilon)$ and λ_k, respectively. This gives us that, if $\hat{\varepsilon} > 0$ is such that

$$\hat{\varepsilon}(\|\check{W}\|_2 + \|\mathbf{r}_k\|_2) < \min\{\gamma_{k-1}, \gamma_k\},$$

then for all ε, such that $|\varepsilon| \leq \hat{\varepsilon}$, we have

$$|\sigma_k - \sigma_k(\varepsilon)| \leq |\varepsilon|\|\mathbf{r}_k\|_2 \qquad \text{and} \qquad |\sigma_k - \sigma_i(\varepsilon)| > \hat{\varepsilon}\|\mathbf{r}_k\|_2, \quad \text{for all } i \neq k.$$

Then we can use Corollary 3.8.1, where we prove that $\|\check{W}\|_2 = \|W\|_2$, to complete the proof. □

In the next theorem we use the result of Proposition 3.8.1 under the assumption that $W \in \mathbb{R}^{n\times p}$ is a general random matrix. After we have stated the theorem, we shall discuss a way of implementing its result on a special class of random matrices, for which we can approximate $\|W\|_2$ numerically.

Theorem 3.8.1. *Let $A \in \mathbb{R}^{n\times p}$ be a deterministic matrix and $W \in \mathbb{R}^{n\times p}$ be a matrix whose entries are random variables. Then, given a* confidence level, *$1-\alpha$, if $\varepsilon_r^* > 0$ is such that*

$$\mathbb{P}\big[\varepsilon_r^*(\|W\|_2 + \|\mathbf{r}_k\|_2) < \min\{\gamma_{k-1}, \gamma_k\}\big] = 1 - \alpha, \tag{3.51}$$

we have

$$\mathbb{P}\big[|\sigma_k - \sigma_k(\varepsilon)| \leq \varepsilon_r^*\|\mathbf{r}_k\|_2 \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon_r^*\big] \geq 1 - \alpha,$$

$$\mathbb{P}\big[|\sigma_k - \sigma_i(\varepsilon)| > \varepsilon_r^*\|\mathbf{r}_k\|_2 \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon_r^* \quad \textit{and all } i \neq k\big] \geq 1 - \alpha$$

and therefore

$$\mathbb{P}\big[\sigma_k(\varepsilon) \neq \sigma_i(\varepsilon) \quad \textit{for all } \varepsilon, \ \textit{s.t. } |\varepsilon| \leq \varepsilon_r^* \quad \textit{and all } i \neq k\big] \geq 1 - \alpha.$$

Proof. Follows from Theorem 3.5.1, applied to the symmetric matrix $\check{A}$ which is perturbed by $\varepsilon\check{W}$ (defined by (3.49) and (3.50)). □

We now discuss a way of implementing the result of Theorem 3.8.1. Firstly, we quote a result from (Johnstone, 2001) which can be used to approximate $\|W\|_2$ numerically. Secondly, we discuss a way of approximating the c.d.f. of $\|W\|_2+\|\mathbf{r}_k\|_2$, which is finally used to find ε_r^* defined by (3.51).

We start by stating the main result of (Johnstone, 2001) assuming the entries of the matrix $W' \in \mathbb{R}^{n\times p}$ are i.i.d. random variables distributed $\mathcal{N}(0,1)$. Then we adapt that result to the class of matrices $W \in \mathbb{R}^{n\times p}$, whose entries are i.i.d. random variables distributed $\mathcal{N}\left(0, \frac{1}{np}\right)$.

Let the eigenvalues of the matrix $W'^T W'$ be denoted by $l_1' > l_2' > \cdots > l_p'$. Let us also define the centre and scaling constants, μ_{np} and σ_{np}, by

$$\mu_{np} := \frac{1}{np}(\sqrt{n-1}+\sqrt{p})^2 \tag{3.52}$$

and

$$\sigma_{np} := \frac{\sqrt{n-1}+\sqrt{p}}{np}\left(\frac{1}{\sqrt{n-1}}+\frac{1}{\sqrt{p}}\right)^{\frac{1}{3}}. \tag{3.53}$$

Finally, let us denote by $F(s)$, as in §2, equation (2.9), the distribution function

$$F(s) = \exp\left(-\frac{1}{2}\left(\int_s^\infty (x-s)q(x)^2dx + \int_s^\infty q(x)dx\right)\right),$$

where q solves the Painlevé II ODE

$$q'' = sq + 2q^3 \tag{3.54}$$

with boundary condition

$$q(s) \sim \mathrm{Ai}(s), \qquad \text{as } s \to \infty, \tag{3.55}$$

where Ai is the Airy function. Under these conditions, if n and p are such that $n/p \to c \geq 1$, as $n \to \infty$, we have

$$\frac{l_1' - np\mu_{np}}{np\sigma_{np}} \xrightarrow{\mathcal{D}} \zeta, \tag{3.56}$$

where $F(s)$ is the c.d.f. of the random variable ζ (see (Johnstone, 2001) for proof and further discussions of this result).

Now, let us consider a matrix $W \in \mathbb{R}^{n\times p}$, whose entries are i.i.d. random variables

distributed $\mathcal{N}\left(0, \frac{1}{np}\right)$. Then $W = \frac{1}{\sqrt{np}} W$ and thus

$$W^T W = \frac{1}{np} W'^T W, \qquad \text{which implies} \quad l_i = \frac{1}{np} l'_i, \quad 1 \leq i \leq n,$$

where $l_1 > l_2 > \cdots > l_p$ are the eigenvalues of $W^T W$. Hence, after multiplying the numerator and the denominator of (3.56) by $1/np$, we obtain

$$\frac{l_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{\mathcal{D}} \zeta,$$

where $F(s)$ is the c.d.f. of the random variable ζ. Thus, the distribution (or the c.d.f.) of the eigenvalue l_1 can be approximated by that of $\mu_{np} + \sigma_{np}\zeta$, where in §2.2 we have given details of how one obtains $F(s)$ by solving (3.54) with the boundary condition (3.55). It is a standard result in Linear Algebra that

$$l_1 = \sigma_1^2,$$

where σ_1 is the largest singular value of W. Hence, one can approximate the c.d.f. of σ_1 by that of $\sqrt{\mu_{np} + \sigma_{np}\zeta}$, which can easily be obtained numerically from $F(s)$, the c.d.f. of ζ. In (Johnstone, 2001) the situation, in which $n \geq p$ and $n/p \to c \leq 1$, is considered by reversing the roles of n and p in (3.52) and (3.53). Thus, we have a way of approximating the c.d.f. of $\|W\|_2$. Numerical tests (not included in this thesis) show that the c.d.f. of $\sqrt{\mu_{np} + \sigma_{np}\zeta}$ approximates that of σ_1 better when the values of n/p is closer to one. When p is significantly smaller than n, one needs to tune the centre and scaling constants, μ_{np} and σ_{np}, to get a better approximation. This is not discussed here further.

Next, in order to approximate the c.d.f. of the sum $\|W\|_2 + \|\mathbf{r}_k\|_2$, we can make the assumption that these two random variables are independent. This is the same as in §3.5, where we used the assumption that $\|B\|_2$ is independent of $\|B(:,k)\|_2$ in order to approximate the c.d.f. of their sum. However, we have not tested the reliability of such an assumption in the rectangular case yet.

As a conclusion, the theory in this section can potentially be implemented in similar way to the theory in §3.5 and, as a consequence, the value of ε_r^* can be approximated using similar techniques.

3.9. Numerical comparisons between ε_1^*, ε_2^* and ε_3^*.

In this section we present some numerical experiments to illustrate the theory in §3.3 – §3.5.

Experiment 3.9.1. *In this experiment we calculate and compare ε_1^*, ε_2^* and ε_3^* for different values of n, $n = 20, 50, 100, 200$ and 5000, using the formulae*

$$\varepsilon_1^* := \frac{\alpha \min\{\gamma_{k-1}, \gamma_k\}}{2\mathbb{E}[\|B\|_2]}, \qquad \varepsilon_2^* := \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2G_n^{-1}(1-\alpha)}, \quad \text{and} \quad \varepsilon_3^* := \frac{\min\{\gamma_{k-1}, \gamma_k\}}{F^{-1}_{\|B\|_2+\|B(:,k)\|_2}(1-\alpha)}, \tag{3.57}$$

where $G_n(t)$ is the c.d.f. of $\|B\|_2$, $1-\alpha = 0.9$ and $\min\{\gamma_{k-1}, \gamma_k\} = 1$. The formulae in (3.57) were given in (3.19), (3.23) and (3.28), respectively.

We also compute $\varepsilon_1^*/\varepsilon_2^*$ and $\varepsilon_2^*/\varepsilon_3^*$ for the values of n given above, in order to compare them with the asymptotic approximations given in (3.24) and in (3.39), respectively.

We briefly recall here that, firstly, $\mathbb{E}[\|B\|_2]$, when B is SGOE matrix, is calculated using Program 3.3.1 (the result of the program, `expect_norm`, has to be further multiplied by $\sqrt{\frac{2}{n}}$). Secondly, the value of $G_n(t)$ is computed via Program 2.3.1 and then it is inverted using built-in MATLAB functions. Thirdly, the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$, $F_{\|B\|_2+\|B(:,k)\|_2}(t)$, is calculated using the assumption that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent. Thus, $F_{\|B\|_2+\|B(:,k)\|_2}(t)$ is obtained as a convolution between the c.d.f.'s of $\|B\|_2$ and $\|B(:,k)\|_2$ (see description of Experiment 3.9.2 for further details) and is also inverted numerically in a standard way.

Results and Discussion. *The results in Table 3.1 show that ε_3^* is greater than ε_1^* and ε_2^* and thus it is a better solution to Problem 3.1.1. Also, from column 1 in the table we can see that the values of ε_1^* approach $0.025 = \frac{\alpha}{4}$, which confirms (experimentally) our still unproved statement that*

$$\mathbb{E}[\|B\|_2] \to 2, \quad \text{as } n \to \infty,$$

when B is SGOE matrix. Further, from column 2 we note that ε_2^* approaches 0.25, also in agreement with

$$G_n^{-1}(1-\alpha) \to 2, \quad \text{as } n \to \infty.$$

Also, column 3 in Table 3.1 shows that ε_3^* approaches $\frac{1}{3}$, fact which agrees with

$$\|B\|_2 + \|B(:,k)\|_2 \xrightarrow{\mathcal{D}} 3 \qquad \text{and thus} \qquad \frac{1}{\|B\|_2 + \|B(:,k)\|_2} \xrightarrow{\mathcal{D}} \frac{1}{3} \quad \text{as } n \to \infty,$$

which is a consequence of Proposition 3.6.2.

Finally, the last two columns of Table 3.1 indicate that

$$\frac{\varepsilon_1^*}{\varepsilon_2^*} \text{ approaches } 0.1 = \alpha \qquad \text{and} \qquad \frac{\varepsilon_2^*}{\varepsilon_3^*} \text{ approaches } \frac{3}{4}, \quad \text{as } n \text{ increases.}$$

This supports the asymptotic results given in (3.24) *and* (3.39).

	ε_1^*	ε_2^*	ε_3^*	$\varepsilon_1^*/\varepsilon_2^*$	$\varepsilon_2^*/\varepsilon_3^*$
$n = 20$	0.0259	0.2348	0.3092	0.1102	0.7593
$n = 50$	0.0255	0.2415	0.3187	0.1054	0.7577
$n = 100$	0.0253	0.2446	0.3233	0.1034	0.7566
$n = 200$	0.0252	0.2466	0.3264	0.1021	0.7554
$n = 5000$	0.0250	0.2496	0.3320	0.1002	0.7518

Table 3.1: *The values of* ε_1^*, ε_2^*, ε_3^*, $\frac{\varepsilon_1^*}{\varepsilon_2^*}$ *and* $\frac{\varepsilon_2^*}{\varepsilon_3^*}$ *for* $n = 20, 50, 100, 200$ *and* 5000.

Experiment 3.9.2. *In this experiment we test the assumption, which we made in §3.5, about the independence of the random variables* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *when* B *is SGOE matrix (see Definition 3.1.1). More precisely, we calculate the c.d.f. of the sum* $\|B\|_2+\|B(:,k)\|_2$*, under the assumption that* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent, and denote that c.d.f. by* $F_n^{(T)}(t)$*, and compare the result with the c.d.f. of* $\|B\|_2+\|B(:,k)\|_2$ *obtained by simulation (i.e. by creating many samples of the SGOE matrix* B *and calculating* $\|B\|_2 + \|B(:,k)\|_2$ *for each of them) which we denote by* $F_n^{(S)}(t)$*.*

We calculate $F_n^{(T)}(t)$ *as a convolution of the c.d.f.'s of* $\|B\|_2$ *and* $\|B(:,k)\|_2$*. Theoretically, if* $f_{\|B\|_2}(t)$ *and* $f_{\|B(:,k)\|_2}$ *are the probability density functions (p.d.f.'s) of* $\|B\|_2$ *and* $\|B(:,k)\|_2$*, respectively, and we assume that both random variables are independent, then the p.d.f. of* $\|B\|_2 + \|B(:,k)\|_2$*,* $f_{\|B\|_2+\|B(:,k)\|_2}(t)$*, is given by*

$$f_{\|B\|_2+\|B(:,k)\|_2}(t) := \int_{-\infty}^{\infty} f_{\|B\|_2}(x) f_{\|B(:,k)\|_2}(t-x)dx$$

and the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$ *is further given by*

$$F_{\|B\|_2+\|B(:,k)\|_2}(t) = \int_{-\infty}^{t} f_{\|B\|_2+\|B(:,k)\|_2}(x)dx.$$

Thus, $F_n^{(T)}(t)$ *is a numerical approximation of* $F_{\|B\|_2+\|B(:,k)\|_2}(t)$*.*

We now describe how we obtain the c.d.f.'s of $\|B\|_2$ *and* $\|B(:,k)\|_2$*. The c.d.f. of* $\|B\|_2$ *is obtained by using results from §2, that is, we calculate it from the solution of*

an initial value problem (c.f. §2.3). For the c.d.f. of $\|B(:,k)\|_2$ *we use the fact that*

$$\|B(:,k)\|_2^2 = B_{kk}^2 + \sum_{i=1, i\neq k}^{n} B_{ik}^2$$

and therefore $\|B(:,k)\|_2 = \xi + \eta$*, where* ξ *and* η *are independent random variables, which satisfy* $n\xi \in \chi^2_{n-1}$ *and* $\frac{n}{2}\eta \in \chi^2_1$*. Hence, the c.d.f. of* $\|B(:,k)\|_2$ *can be obtained by standard built-in MATLAB functions.*

The results from this experiment, for $n = 20, 50, 100$ *and* 200*, are presented in Figures 3-1, 3-2, 3-3 and 3-4.*

Results and Discussion. *We can see from Figures 3-1, 3-2, 3-3 and 3-4 that for* $n = 50$ *the agreement between* $F_n^{(S)}(t)$ *and* $F_n^{(T)}(t)$*is satisfactory and for* $n = 100$ *and* 200 *the difference between both c.d.f.'s is negligible. Therefore, the results from this experiment suggest that the assumption of the independence of the random variables* $\|B\|_2$ *and* $\|B\|_2$ *is fairly accurate for values of* n *not less than* 50 *and the accuracy improves for larger values of* n*.*

This experiment also indicates that $\|B\|_2 + \|B(:,k)\|_2 \xrightarrow{\mathcal{D}} 3$*, as* n *increases, a result which can be theoretically confirmed by applying Slutsky's Theorem (c.f. Theorem A.1.7) to the limits of* $\|B\|_2$ *and* $\|B(:,k)\|_2$*, found in §3.6.*

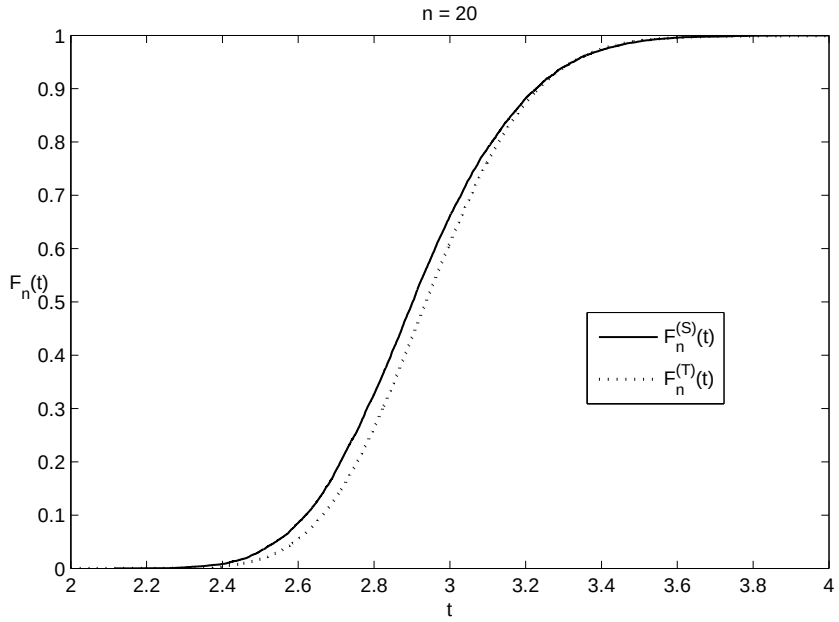

Figure 3-1: *Comparison between the c.d.f.'s of* $\|B\|_2 + \|B(:,k)\|_2$*, obtained by simulation and by theory, assuming* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent. The former is denoted by* $F_n^{(S)}(t)$ *and the latter, by* $F_n^{(T)}(t)$*. Here* $n = 20$*.*

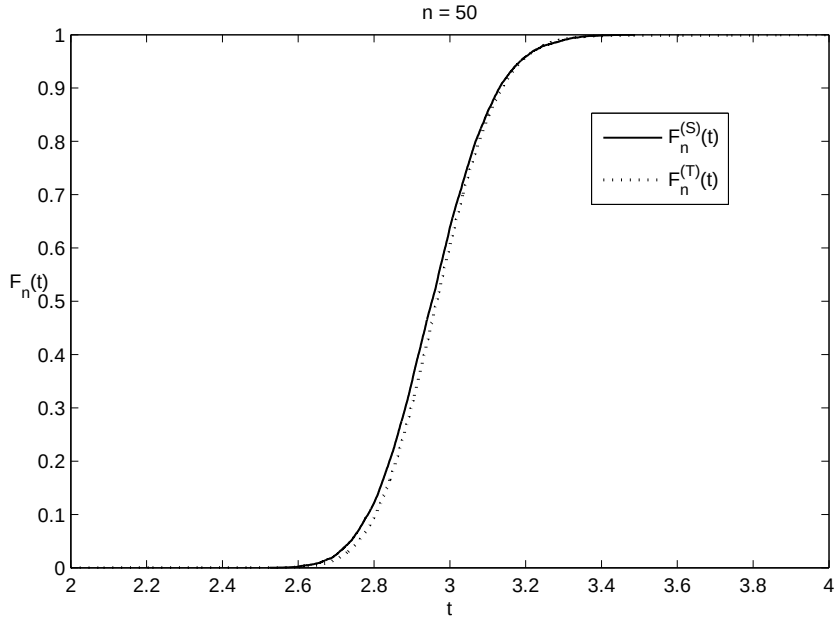

Figure 3-2: *Comparison between the c.d.f.'s of* $\|B\|_2 + \|B(:,k)\|_2$*, obtained by simulation and by theory, assuming* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent. The former is denoted by* $F_n^{(S)}(t)$ *and the latter, by* $F_n^{(T)}(t)$*. Here* $n = 50$*.*

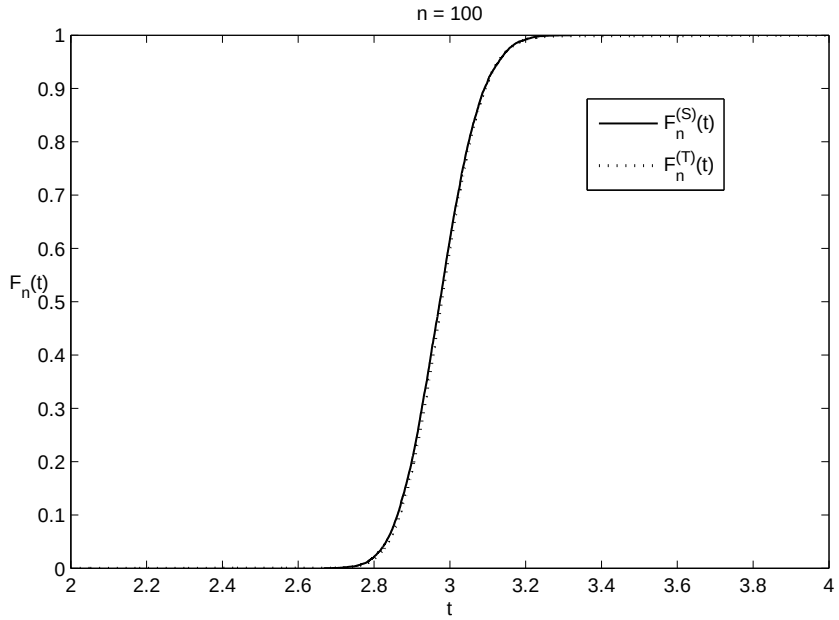

Figure 3-3: *Comparison between the c.d.f.'s of* $\|B\|_2 + \|B(:,k)\|_2$*, obtained by simulation and by theory, assuming* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent. The former is denoted by* $F_n^{(S)}(t)$ *and the latter, by* $F_n^{(T)}(t)$*. Here* $n = 100$*.*

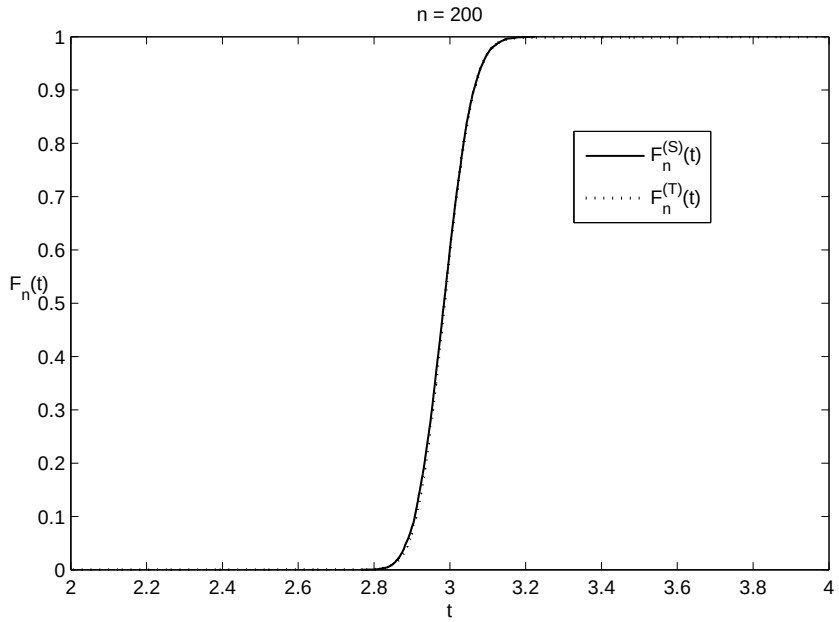

Figure 3-4: *Comparison between the c.d.f.'s of* $\|B\|_2 + \|B(:,k)\|_2$*, obtained by simulation and by theory, assuming* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *are independent. The former is denoted by* $F_n^{(S)}(t)$ *and the latter, by* $F_n^{(T)}(t)$*. Here* $n = 200$*.*

Experiment 3.9.3. *In this experiment we test the theoretical approach of finding* ε_3^**, suggested in §3.5, when the independence of* $\|B\|_2$ *and* $\|B(:,k)\|_2$ *is assumed. From Theorem 3.5.1 we have that, when* ε *is such that* $|\varepsilon| \leq \varepsilon_3^*$*, the following inequality is satisfied*

$$|\lambda_k - \lambda_k(\varepsilon)| \leq |\varepsilon| \|B(:,k)\|_2$$

with probability not less than $1-\alpha$*. So, the goal of this experiment is also to check the magnitude of the difference between*

$$|\lambda_n - \lambda_n(\varepsilon_3^*)| \qquad \textit{and} \qquad |\varepsilon_3^*| \|B(:,n)\|_2$$

and the amount by which

$$\mathbb{P}[|\lambda_k - \lambda_k(\varepsilon_3^*)| \leq |\varepsilon_3^*| \|B(:,k)\|_2] \qquad \textit{exceeds} \qquad 1-\alpha.$$

The tests we do here are for $n = 20, 50, 100$ *and* 200*. The* confidence level, $1-\alpha$*, is set at* 0.9 *and the index* k *is chosen equal to* n*. The matrix to be perturbed,* A*, is a diagonal matrix whose entries on the main diagonal are simulated for each* n *as* $n-1$ *independent, uniformly distributed random variables over the interval* $(0,1)$*. The largest entry on the main diagonal of* A *is fixed at* 1.5 *for all values of* n *considered*

here, in order to keep the gap between two largest entries of A approximately equal to 0.5*, and thus relatively independent of* n*.*

The experiments consists of the following: We find the cumulative distribution function (c.d.f.) of $\|B\|_2 + \|B(:,n)\|_2$ *under the assumption that both random variables entering the sum are independent, which we tested in Experiment 3.9.2. Thus, the c.d.f. of* $\|B\|_2 + \|B(:,n)\|_2$*,* $F_{\|B\|_2+\|B(:,n)\|_2}(t)$*, is found as the convolution between the c.d.f.'s of* $\|B\|_2$ *and* $\|B(:,n)\|_2$*. (In Experiment 3.9.2 we described the way in which we obtain both c.d.f.'s.) We find* ε_3^* *by inverting the c.d.f. of* $\|B\|_2 + \|B(:,n)\|_2$*, that is, we let*

$$\varepsilon_3^* := \frac{\gamma_{n-1}}{F^{-1}_{\|B\|_2+\|B(:,n)\|_2}(1-\alpha)},$$

where $\gamma_{n-1} = \lambda_n - \lambda_{n-1} \approx 0.5$*.*

Once we have obtained ε_3^**, we simulate the difference* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$ $10\,000$ *times, by simulating the SGOE matrix* B*. From the simulations of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$ *we obtain its empirical c.d.f. and compare it with the c.d.f. of* $|\varepsilon_3^*|\|B(:,n)\|_2$*. Both c.d.f.'s are plotted in Figures 3-5, 3-6, 3-7 and 3-8, together with the value of* ε_3^**. The* x*-axes of Figures 3-5, 3-6, 3-7 and 3-8 are set so that their lengths are equal to* γ_{n-1}*.*

Results and Discussion. *In Figures 3-5, 3-6, 3-7 and 3-8 we can see that the difference between the c.d.f.'s of* $|\varepsilon_3^*|\|B(:,n)\|_2$ *and* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$ *increases, as* n *increases and for* $n = 200$ *the magnitude of* $|\varepsilon_3^*|\|B(:,n)\|_2$ *is more than twice as large as that of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$*. This indicates that, although* ε_3^* *is a better solution to Problem 3.1.1 than* ε_1^* *and* ε_2^**, since* $\varepsilon_3^* > \varepsilon_2^*$ *(see the discussion at the end of §3.6) and* $\varepsilon_2^* > \varepsilon_1^*$*, it is still an underestimate of the largest possible solution of Problem 3.1.1. We think the main reason for this is that the Bauer-Fike's Theorem overestimates the difference* $|\lambda_n - \lambda_n(\varepsilon)|$ *by a very large margin.*

Since the length of the x*-axes of Figures 3-5, 3-6, 3-7 and 3-8 is equal to the value of* γ_{n-1} *for each of the tests, we can see by the position of* ε_3^* *on the* x*-axis that*

$$\varepsilon_3^* \to \frac{1}{3}\gamma_{n-1},$$

as n *increases, which confirms the theoretical result following Proposition 3.6.2.*

Also, as n *increases, Figures 3-5, 3-6, 3-7 and 3-8 suggest that*

$$|\varepsilon_3^*|\|B(:,n)\|_2 \xrightarrow{\mathcal{D}} |\varepsilon_3^*|,$$

which implies that $\|B(:,n)\|_2 \xrightarrow{\mathcal{D}} 1$*, as shown in Lemma 3.6.2.*

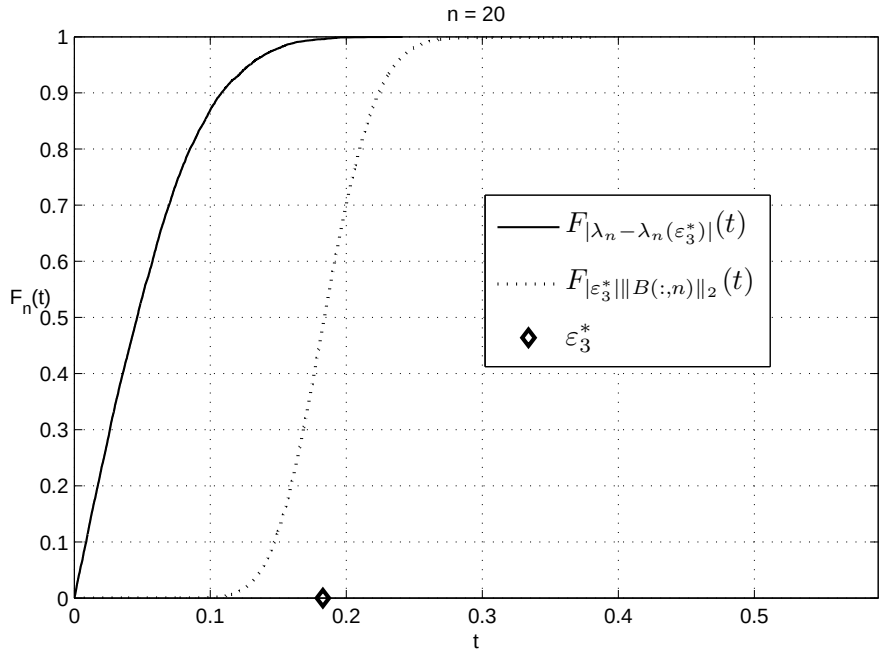

Figure 3-5: *Comparison between the c.d.f. of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$*, obtained by simulation, and that of* $|\varepsilon_3^*|\|B(:,n)\|_2$*. The former is denoted by* $F_{|\lambda_n-\lambda_n(\varepsilon_3^*)|}(t)$ *and the latter, by* $F_{|\varepsilon_3^*|\|B(:,n)\|_2}(t)$*. Here* $n = 20$*.*

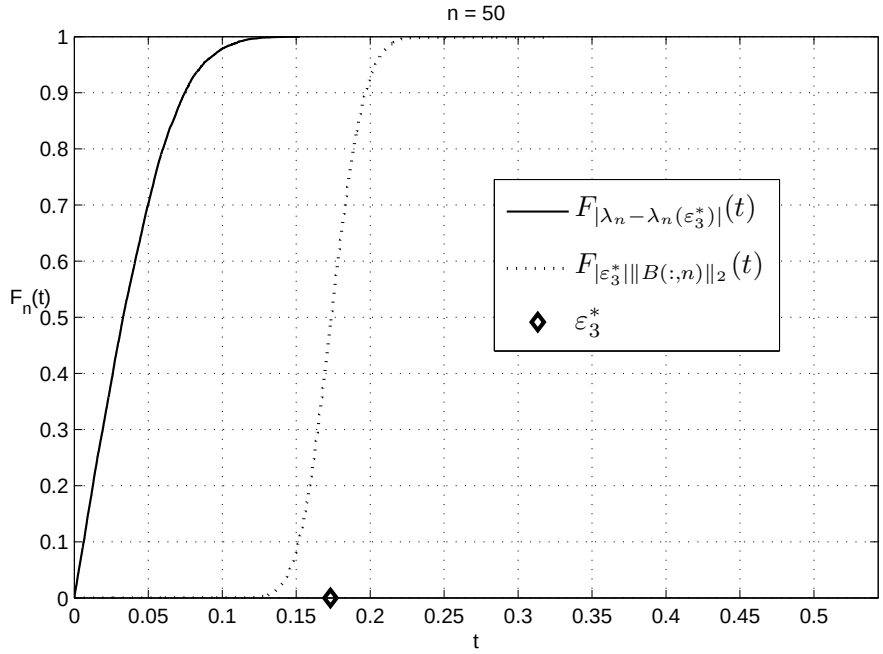

Figure 3-6: *Comparison between the c.d.f. of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$*, obtained by simulation, and that of* $|\varepsilon_3^*|\|B(:,n)\|_2$*. The former is denoted by* $F_{|\lambda_n-\lambda_n(\varepsilon_3^*)|}(t)$ *and the latter, by* $F_{|\varepsilon_3^*|\|B(:,n)\|_2}(t)$*. Here* $n = 50$*.*

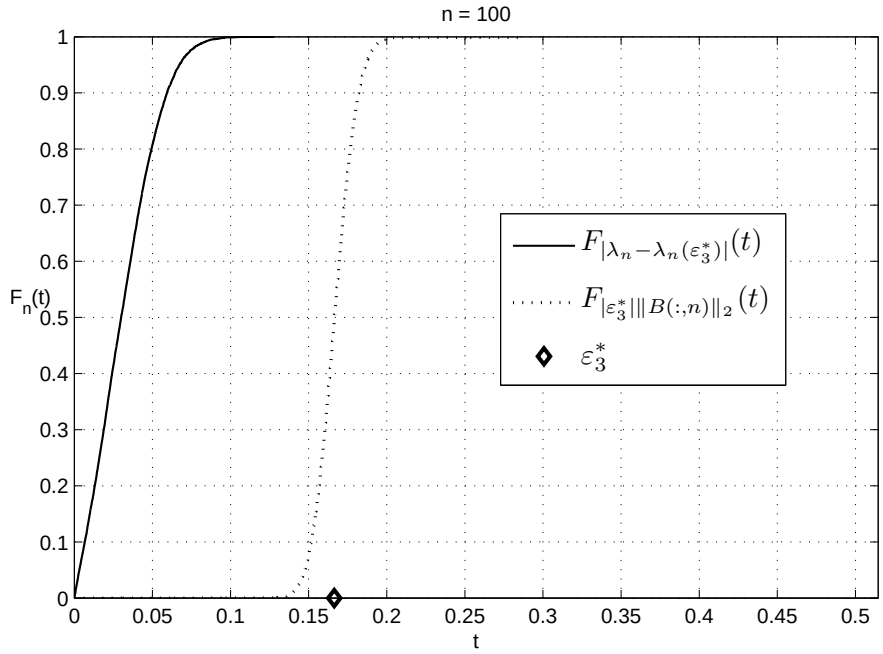

Figure 3-7: *Comparison between the c.d.f. of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$*, obtained by simulation, and that of* $|\varepsilon_3^*|\|B(:,n)\|_2$*. The former is denoted by* $F_{|\lambda_n-\lambda_n(\varepsilon_3^*)|}(t)$ *and the latter, by* $F_{|\varepsilon_3^*|\|B(:,n)\|_2}(t)$*. Here* $n = 100$*.*

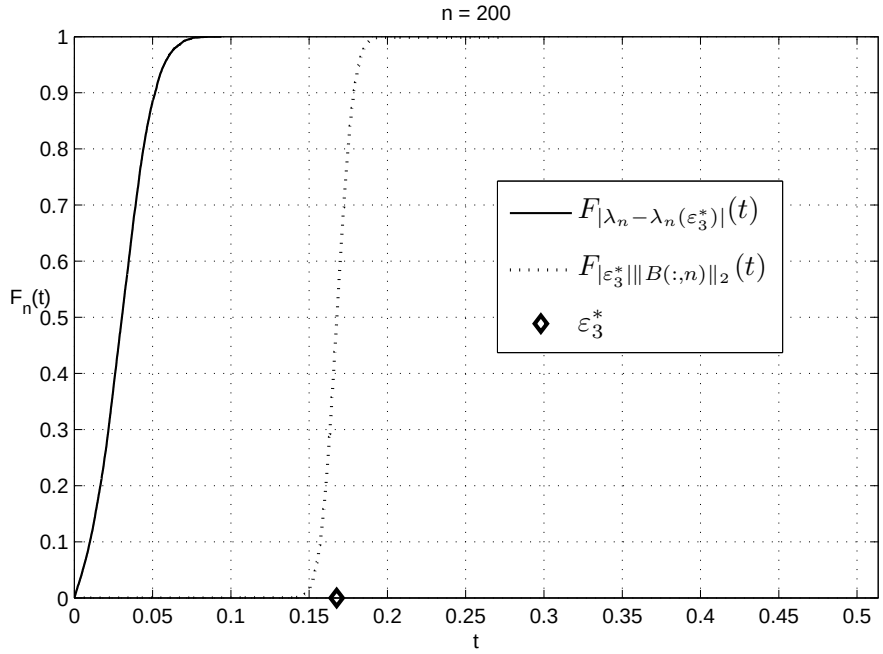

Figure 3-8: *Comparison between the c.d.f. of* $|\lambda_n - \lambda_n(\varepsilon_3^*)|$*, obtained by simulation, and that of* $|\varepsilon_3^*|\|B(:,n)\|_2$*. The former is denoted by* $F_{|\lambda_n-\lambda_n(\varepsilon_3^*)|}(t)$ *and the latter, by* $F_{|\varepsilon_3^*|\|B(:,n)\|_2}(t)$*. Here* $n = 200$*.*

Chapter 4. Perturbation theory using linearisation.

4.1. Introduction.

In this chapter we consider the inverse of Problem 3.1.1 from §3. Given a magnitude of the perturbation, ε_0, of the entries of the matrix A, we bound from above the probability of a swap between a certain eigenvalue and the rest of the eigenvalues in the spectrum of $A(\varepsilon_0) := A + \varepsilon_0 B$, where B is SGOE matrix (see Definition 4.1.1).

The motivation for this problem comes from situations where the magnitude of perturbation is approximately known. For example, in micro-array analysis the magnitude of the errors is known roughly and so analysis of the type in this chapter may provide bounds on clustering using spectral techniques. Since spectral clustering is done by examining the entries of the eigenvector, associated with a certain eigenvalue (or eigenspace associated with a set of eigenvalues), the question of reliability of spectral clustering can be stated as a problem of the stability to perturbation of the part of the spectrum, responsible for the clustering, given the size of the errors. Here we consider only the stability to perturbation of a single eigenvalue. This analysis may be helpful when the spectral clustering is done by the entries of only one eigenvector of the matrix, associated with the network, but we do not consider this further in this thesis.

Next, we recall a definition and provide a mathematical statement of the problem described above. We start by recalling the definition of SGOE matrix from §3.

Definition 4.1.1. We say that the random symmetric matrix B is Scaled GOE matrix, or shortly SGOE matrix, if its entries above and on the main diagonal are independent random variables satisfying

$$B_{ii} \in \mathcal{N}\left(0, \frac{2}{n}\right) \quad \text{for } 1 \leq i \leq n \qquad \text{and} \qquad B_{ij} \in \mathcal{N}\left(0, \frac{1}{n}\right) \quad \text{for } 1 \leq i < j \leq n.$$

The following is the statement of the problem we are concerned with in this chapter.

Problem 4.1.1. *Let $A \in \mathbb{R}^{n\times n}$ be a deterministic symmetric matrix and $B \in \mathbb{R}^{n\times n}$ be SGOE matrix. Further, let $A(\varepsilon) := A + \varepsilon B$ and let also $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $\lambda_1(\varepsilon) \leq \lambda_2(\varepsilon) \leq \cdots \leq \lambda_n(\varepsilon)$ be the eigenvalues of A and $A(\varepsilon)$, respectively. Finally, let λ_k be a simple eigenvalue of A for some $1 \leq k \leq n$. Given an $\varepsilon_0 > 0$, find an upper bound on the probability*

$$\mathbb{P}\big[\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\big]. \tag{4.1}$$

Remark 4.1.1. *In the statement of Problem 4.1.1 the event*

$$\{\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\}$$

means that $\lambda_k(\varepsilon)$ becomes a multiple eigenvalue of $A(\varepsilon)$ for some ε satisfying $|\varepsilon| \leq \varepsilon_0$. Here we have used the fact that the eigenvalues of $A(\varepsilon)$ are continuous functions of ε. Therefore if, for example, the k-th smallest eigenvalue of $A(\varepsilon_0)$ has swapped with $A(\varepsilon_0)$'s $(k+1)$-th smallest eigenvalue, then there will be an ε, such that $0 < \varepsilon < \varepsilon_0$, for which both eigenvalues were equal. Therefore, we have the following relation between events

$$\begin{aligned} &\{\exists j, j \neq k \quad s.t. \quad \lambda_k(\varepsilon_0) \text{ has swapped with } \lambda_j(\varepsilon_0)\} \\ &\qquad \subset \{\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\}. \end{aligned} \tag{4.2}$$

Hence, the quantity which bounds the probability

$$\mathbb{P}[\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)]$$

from above, may serve as a bound from above to the probability

$$\mathbb{P}[\exists j, j \neq k \quad s.t. \quad \lambda_k(\varepsilon_0) \text{ has swapped with } \lambda_j(\varepsilon_0)].$$

In theory, it is not clear whether the inclusion

$$\begin{aligned} &\{\exists j, j \neq k \quad s.t. \quad \lambda_k(\varepsilon_0) \text{ has swapped with } \lambda_j(\varepsilon_0)\} \\ &\qquad \supset \{\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\} \end{aligned}$$

also holds (or at least we are not aware of any such results). However, we can extend the definition of a swap between $\lambda_k(\varepsilon)$ and the rest of the spectrum of $A(\varepsilon)$ by defining

$$\{\lambda_k(\varepsilon_0) \text{ has swapped }\} := \{\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad s.t. \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\}.$$

One practical reason for such an extension is that the event in (4.1) *implies that for some* ε *we can no longer be using the eigenvector* $\mathbf{v}_k(\varepsilon)$ *in an unique way, but rather we get input from vectors which are orthogonal to* $\mathbf{v}_k(\varepsilon)$*. In clustering of networks this would mean that if we were using the eigenvector corresponding to the* k*-th smallest eigenvalue of* A *for clustering, then clustering with respect to* $\mathbf{v}_k$*, corresponding to the "original data", and clustering with respect to* $\mathbf{v}_k(\varepsilon)$*, corresponding to "perturbed data", would lead to different results, due to the contribution of the directions orthogonal to* $\mathbf{v}_k(\varepsilon)$*. In other words, even though the event in* (4.1) *might not mean that* $\lambda_k(\varepsilon_0)$ *has swapped according to* (4.2)*, we find it reasonable to extend the definition of a swap of* $\lambda_k(\varepsilon_0)$*, by making it equivalent to the event entering* (4.1) *(see Definition 4.1.2 below).*

Since ε_0 *is the magnitude of the perturbation, we may have positive as well as negative perturbations. Thus, we require* $|\varepsilon| \leq \varepsilon_0$ *in* (4.1)*, instead of only* $0 \leq \varepsilon \leq \varepsilon_0$*.*

Definition 4.1.2. In the settings of Problem (4.1.1) let $\varepsilon_0 > 0$ be given. Then we define the event

$$\{\lambda_k(\varepsilon_0) \text{ has swapped } \} := \{\exists \varepsilon, |\varepsilon| \leq \varepsilon_0, \exists j, j \neq k, \quad \text{s.t.} \quad \lambda_k(\varepsilon) = \lambda_j(\varepsilon)\}. \tag{4.3}$$

The gaps between the consecutive eigenvalues of A will be denoted by γ_i, that is,

$$\gamma_i := \lambda_{i+1} - \lambda_i, \qquad 1 \leq i \leq n-1, \tag{4.4}$$

and the gaps between the eigenvalues of $A(\varepsilon)$ will be denoted by

$$\gamma_i(\varepsilon) := \lambda_{i+1}(\varepsilon) - \lambda_i(\varepsilon), \qquad 1 \leq i \leq n-1. \tag{4.5}$$

The plan of this chapter is as follows. Firstly, in §4.2 we briefly revise the linear approximation of perturbed simple eigenvalues of symmetric matrices and using this, we provide an upper bound on the probability $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$, assuming ε_0 is given. We test the theory in this section and discuss its limitations in §4.4. Secondly, in §4.3 we combine the Bauer-Fike Theorem (Corollary 3.2.1 in §3.2) with Theorem 3.2.2 from §3.2 to state a result similar to Theorem 3.5.1 in §3.5. The result in §4.3 provides an upper bound on $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$, which is crude for the moment. The advantages and disadvantages of this approach are discussed at the end of §4.3. Finally, in §4.4 we provide a numerical experiment which compares the upper bounds from §4.2 with $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$ obtained by simulation.

4.2. Bounding the probability of a swap from above by linearisation.

In this section we use first-order Perturbation Theory for symmetric matrices, in the case when the matrix by which we perturb is random, to provide an upper bound on the probability $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$. A very detailed account of the first-order Stochastic Perturbation Theory to a more general class of matrices can be found in (Stewart, 1990). However, the problem we consider here (Problem 4.1.1) is somewhat different. We solve our problem, using the fact that the distributions of the linear approximations to the perturbed eigenvalues are very easy to work with. In our case, when B is SGOE matrix, these linear approximations are normally distributed random variables. We use this extensively in Proposition 4.2.1 to provide a nice formula, (4.21), for the calculation of the probability on the left hand side of (4.12). The latter, as we show in the discussion following Remark 4.2.1, can be used as an "approximate" upper bound on $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$. The results of this section are tested numerically in Experiment 4.4.1.

We present our main result here, (4.12), in the case when B is SGOE matrix. However, result similar to (4.12) can be proved for any random symmetric matrix, B, as long as one could provide formula, analogical to (4.21) in Proposition 4.2.1, for the calculation of the upper bound in that result.

Finally, the results in this section assume that the matrix which is perturbed has a simple spectrum, but this restriction can easily be relaxed by requiring only the eigenvalue we are interested in, λ_k, to be simple.

Let the eigenvalues of A be simple, that is, let $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, and also let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be their corresponding eigenvectors of unit length. Then we know from first-order Perturbation Theory of (deterministic) symmetric matrices that the eigenvalues of $A(\varepsilon)$, $\lambda_i(\varepsilon)$, are given by

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon \mathbf{v}_i^T B \mathbf{v}_i + \mathcal{O}(\varepsilon^2), \qquad 1 \leq i \leq n. \tag{4.6}$$

In this section we shall use (4.6) neglecting higher-order order terms. In other words, we shall use the linear approximation, $\tilde{\lambda}_i(\varepsilon)$, instead of $\lambda_i(\varepsilon)$ itself, where

$$\tilde{\lambda}_i(\varepsilon) = \lambda_i + \varepsilon \mathbf{v}_i^T B \mathbf{v}_i, \qquad 1 \leq i \leq n. \tag{4.7}$$

Essentially, this means that we no longer work with $\lambda_i(\varepsilon)$, $1 \leq i \leq n$, whose exact distributions are hard to obtain analytically and also difficult for the manipulations needed here. Instead, we work entirely with $\tilde{\lambda}_i(\varepsilon)$, $1 \leq i \leq n$, whose distributions

don't have such disadvantages, in the hope that $\tilde{\lambda}_i(\varepsilon)$ will give us an idea of the real behaviour of $\lambda_i(\varepsilon)$. Despite the neat results we obtain using $\tilde{\lambda}_i(\varepsilon)$, $1 \leq i \leq n$, instead of $\lambda_i(\varepsilon)$, $1 \leq i \leq n$, we prove later that such an approach has its limitations. This is confirmed numerically in Experiment 4.4.1 (in §4.4). Roughly speaking, one of the limitations of using first-order approximations to $\lambda_i(\varepsilon)$ consists of the fact that the linear term, $\varepsilon \mathbf{v}_i^T B \mathbf{v}_i$, converges to zero when ε and $\|B\|_2$ are kept relatively unchanged (e.g. when B is SGOE matrix), while the dimension $n \to \infty$. To the best of our knowledge, this problem has received very little (if any) attention in the literature. If we consider second-order expansions:

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon \mathbf{v}_i^T B \mathbf{v}_i + \varepsilon^2 \sum_{j=1, j\neq i}^{n} \frac{(\mathbf{v}_k^T B \mathbf{v}_i)^2}{\lambda_i - \lambda_j} + \mathcal{O}(\varepsilon^3),$$

it can be shown that the second-order term, $\varepsilon^2 \sum_{j=1, j\neq i}^{n} \frac{(\mathbf{v}_k^T B \mathbf{v}_i)^2}{\lambda_i - \lambda_j}$, doesn't converge to zero (as $n \to \infty$) if ε is a constant and B is SGOE matrix (i.e. $\|B\|_2 \xrightarrow{\mathcal{D}} 2$). However, even in this case, because of the presence of the multipliers $\frac{1}{\lambda_i - \lambda_j}$, $j \neq i$, in the second-order term, great care should be taken for the quantity $\min_{j, j\neq i} |\lambda_j - \lambda_i|$ to be bounded away from zero (as $n \to \infty$) or at least its convergence to zero to be "slow enough". Otherwise, if the convergence $\min_{j, j\neq i} |\lambda_j - \lambda_i| \to 0$ is "too fast", the value of the second-order term may become unbounded as $n \to \infty$. When the term $\min_{j, j\neq i} |\lambda_j - \lambda_i|$ is bounded away from zero, it can easily be shown that in some cases the second-order term converges to a constant in distribution. This means that

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon c_1 n^{-1/2} + \varepsilon^2 c_2 + \mathcal{O}(\varepsilon^3), \tag{4.8}$$

where c_1 and c_2 are random variables whose distributions don't depend on n. Here we have used the fact that the coefficient in front of ε in the first-order term, $\mathbf{v}_i^T B \mathbf{v}_i \in \mathcal{N}\left(0, \frac{2}{n}\right)$ and thus $n^{1/2} \mathbf{v}_i^T B \mathbf{v}_i \in \mathcal{N}(0, 2)$, which doesn't depend on n. Therefore, roughly speaking, in order for the first-order coefficient in (4.8) to be more significant than the second-order coefficient, we need

$$\varepsilon^2 c_2 \ll \varepsilon c_1 n^{-1/2},$$

which leads us to the requirement $\varepsilon \ll c n^{-1/2}$, where c is a random variable whose distribution doesn't depend on n.

The analysis of second-order expansion of $\lambda_i(\varepsilon)$ is still work in progress and thus, it is not presented in this chapter. Our aim here is only to show the limitation of the first-order theory and hence, to suggest that second-order theory might be a remedy.

Now, we shall use $\tilde{\lambda}_i(\varepsilon)$, given in (4.7), assuming that B is SGOE matrix. In doing this we use the discussion preceding Theorem 3.5.1 in §3.5, where we showed that, without loss of generality, we may assume that

$$A(\varepsilon) = \Lambda + \varepsilon B,$$

where $A = V\Lambda V^T$ is the spectral decomposition of A and B is SGOE matrix. Therefore (4.7) becomes

$$\tilde{\lambda}_i(\varepsilon) = \lambda_i + \varepsilon B_{ii}, \qquad 1 \leq i \leq n.$$

Note, we show in §2.5 that the distribution of $\mathbf{v}_i^T B \mathbf{v}_i$ is the same as that of B_{ii}. Thus, $\tilde{\lambda}_i(\varepsilon)$'s distribution doesn't depend on whether we work with the matrix A, or with the diagonal matrix Λ, whose diagonal entries are A's eigenvalues – in both cases $\tilde{\lambda}_i(\varepsilon)$ has the same distribution.

We recall that the diagonal elements of B, B_{ii}, $1 \leq i \leq n$, are distributed $\mathcal{N}\left(0, \frac{2}{n}\right)$. Therefore the random variables $\tilde{\lambda}_i(\varepsilon)$, $1 \leq i \leq n$, are distributed $\mathcal{N}\left(\lambda_i, \frac{2}{n}\right)$, which makes them easy for manipulation.

Let us define

$$\tilde{\gamma}_i(\varepsilon) := \tilde{\lambda}_{i+1}(\varepsilon) - \tilde{\lambda}_i(\varepsilon) = \gamma_i + \varepsilon(B_{i+1,i+1} - B_{ii}), \qquad 1 \leq i \leq n,$$

where γ_i was defined in (4.4). In fact, the random variables $\tilde{\gamma}_i(\varepsilon)$ are linear approximations to the gaps $\gamma_i(\varepsilon)$, defined in (4.5). This shall be used in Remark 4.2.1, in order to provide a link between the events in (4.14). More generally,

$$\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon) = \lambda_i - \lambda_j + \varepsilon(B_{ii} - B_{jj})$$

and thus one can prove that

$$|\lambda_i - \lambda_j| - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} \leq |\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| \leq |\lambda_i - \lambda_j| + 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\},$$

which implies

$$\min_{j,j\neq i}\{|\lambda_i - \lambda_j|\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} \leq \min_{j,j\neq i}|\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| \leq \min_{j,j\neq i}\{|\lambda_i - \lambda_j|\} + 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\}.$$

The latter is equivalent to

$$\min\{\gamma_{i-1}, \gamma_i\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} \leq \min_{j,j\neq i}|\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| \leq \min\{\gamma_{i-1}, \gamma_i\} + 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} \tag{4.9}$$

for all $1 \leq i \leq n$ almost surely. In order for the term $\min\{\gamma_{i-1}, \gamma_i\}$ to be defined properly for $i = 1$, we may for example let $\gamma_0 := +\infty$.

Working with left-hand side of the last inequality we have

$$\min\{\gamma_{i-1}, \gamma_i\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} \leq \min_{j,j\neq i} |\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)|, \quad 1 \leq i \leq n,$$

almost surely. Therefore the following relation between events is satisfied:

$$\left\{\min\{\gamma_{i-1}, \gamma_i\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} > 0\right\} \subset \{\min_{j,j\neq i} |\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| > 0\} \tag{4.10}$$

for all $\varepsilon \in \mathbb{R}$, where i is some index satisfying $1 \leq i \leq n$. This implies

$$\mathbb{P}\left[|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} < \frac{\min\{\gamma_{i-1}, \gamma_i\}}{2}\right] \leq \mathbb{P}[\min_{j,j\neq i} |\tilde{\lambda}_i(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| > 0] \tag{4.11}$$

for all $\varepsilon \in \mathbb{R}$. Therefore, we can prove the following theorem.

Theorem 4.2.1. *In the settings of this section the following inequality holds*

$$\mathbb{P}\left[\max_{1\leq j\leq n} |B_{jj}| \geq \frac{\min\{\gamma_{k-1}, \gamma_k\}}{2\varepsilon_0}\right] \geq 1 - \mathbb{P}\big[\min_{j,j\neq k} |\tilde{\lambda}_k(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| > 0 \text{ for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_0\}\big] \tag{4.12}$$

for any $\varepsilon_0 > 0$.

Proof. Let us assume now that $\varepsilon_0 > 0$ is given. Then from (4.10) we have

$$\begin{aligned}
& \{\min\{\gamma_{k-1}, \gamma_k\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} > 0, \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_0\} \\
= & \bigcap_{\varepsilon, |\varepsilon|\leq\varepsilon_0} \{\min\{\gamma_{k-1}, \gamma_k\} - 2|\varepsilon| \max_{1\leq j\leq n}\{|B_{jj}|\} > 0\} \\
= & \{\min\{\gamma_{k-1}, \gamma_k\} - 2|\varepsilon_0| \max_{1\leq j\leq n}\{|B_{jj}|\} > 0\} \\
\subset & \{\min_{j,j\neq k} |\tilde{\lambda}_k(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| > 0, \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_0\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
& \mathbb{P}[\min\{\gamma_{k-1}, \gamma_k\} - 2|\varepsilon_0| \max_{1\leq j\leq n}\{|B_{jj}|\} > 0] \\
& \qquad \leq \mathbb{P}[\min_{j,j\neq k} |\tilde{\lambda}_k(\varepsilon) - \tilde{\lambda}_j(\varepsilon)| > 0, \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_0],
\end{aligned} \tag{4.13}$$

where the last inequality holds for all $\varepsilon_0 > 0$. Now (4.13) easily leads to (4.12). □

The following remark links the probabilities of the events

$$\{\min_{j,j\neq k}|\tilde{\lambda}_k(\varepsilon)-\tilde{\lambda}_j(\varepsilon)|>0,\quad \text{for all } \varepsilon,\ \text{s.t. } |\varepsilon|\leq\varepsilon_0\}\qquad\text{and}\qquad\{\lambda_k(\varepsilon_0)\text{ has swapped }\}. \tag{4.14}$$

Remark 4.2.1. *Let $\varepsilon_0>0$ be given. For the complement of the event $\{\lambda_k(\varepsilon_0)$ has swapped $\}$, which we denote by $\overline{\{\lambda_k(\varepsilon_0)\text{ has swapped }\}}$, we have*

$$\overline{\{\lambda_k(\varepsilon_0)\text{ has swapped }\}}=\big\{\min\{\gamma_{k-1}(\varepsilon),\gamma_k(\varepsilon)\}>0,\quad \text{for all } \varepsilon,\ \text{s.t. } |\varepsilon|\leq\varepsilon_0\big\},$$

where we have used the continuity of the eigenvalues of $A(\varepsilon)$ as functions of ε. Therefore,

$$\mathbb{P}\big[\min\{\gamma_{k-1}(\varepsilon),\gamma_k(\varepsilon)\}>0,\quad \text{for all } \varepsilon,\ \text{s.t. } |\varepsilon|\leq\varepsilon_0\big]=1-\mathbb{P}\big[\lambda_k(\varepsilon_0)\text{ has swapped }\big].$$

When ε is small we have $\tilde{\lambda}_i(\varepsilon)\approx\lambda_i(\varepsilon)$, $1\leq i\leq n$, and therefore, in particular, $\lambda_i(\varepsilon)-\lambda_j(\varepsilon)\approx\tilde{\lambda}_i(\varepsilon)-\tilde{\lambda}_j(\varepsilon)$, which leads to

$$\min\{\gamma_{k-1}(\varepsilon),\gamma_k(\varepsilon)\}=\min_{j,j\neq k}|\lambda_k(\varepsilon)-\lambda_j(\varepsilon)|\approx\min_{j,j\neq k}|\tilde{\lambda}_k(\varepsilon)-\tilde{\lambda}_j(\varepsilon)|. \tag{4.15}$$

But note that since $\tilde{\lambda}_i(\varepsilon)$ is only a first-order approximation to $\lambda_i(\varepsilon)$, we no longer have the ordering $\tilde{\lambda}_1(\varepsilon)\leq\tilde{\lambda}_2(\varepsilon)\leq\cdots\leq\tilde{\lambda}_n(\varepsilon)$. This is why we have to take

$$\min_{j,j\neq k}|\tilde{\lambda}_k(\varepsilon)-\tilde{\lambda}_j(\varepsilon)|\qquad\textit{instead of only}\qquad\min\{\tilde{\gamma}_{k-1}(\varepsilon),\tilde{\gamma}_k(\varepsilon)\}.$$

Hence, from (4.15) *we obtain*

$$\begin{aligned}\mathbb{P}[\min_{j,j\neq k}|\tilde{\lambda}_k(\varepsilon)-\tilde{\lambda}_j(\varepsilon)|>0,\quad &\textit{for all } \varepsilon,\ \textit{s.t. } |\varepsilon|\leq\varepsilon_0] \\ &\approx\mathbb{P}[\min\{\gamma_{k-1}(\varepsilon),\gamma_k(\varepsilon)\}>0,\quad \textit{for all } \varepsilon,\ \textit{s.t. } |\varepsilon|\leq\varepsilon_0] \\ &=1-\mathbb{P}[\lambda_k(\varepsilon_0)\textit{ has swapped }],\end{aligned} \tag{4.16}$$

when ε_0 is small.

The conclusion from Remark 4.2.1 is that

$$1-\mathbb{P}[\min_{j,j\neq k}|\tilde{\lambda}_k(\varepsilon)-\tilde{\lambda}_j(\varepsilon)|>0,\quad \text{for all } \varepsilon,\ \text{s.t. } |\varepsilon|\leq\varepsilon_0] \tag{4.17}$$

can be used as an approximate upper bound on the probability $\mathbb{P}[\lambda_k(\varepsilon)$ has swapped $]$.

Thus, from Theorem 4.2.1, the probability

$$\mathbb{P}\left[\max_{1\leq j\leq n}|B_{jj}|\geq\frac{\min\{\gamma_{k-1},\gamma_k\}}{2\varepsilon_0}\right] \tag{4.18}$$

can be used as an approximate upper bound on the probability

$$\mathbb{P}[\lambda_k(\varepsilon_0)\text{ has swapped }]. \tag{4.19}$$

The following proposition gives a way of computing the probability in (4.18).

Proposition 4.2.1. *Let us denote* $\eta_n := \max_{1\leq i\leq n}\{|B_{ii}|\}$*, where* B_{ii}*,* $1\leq i\leq n$*, are independent identically distributed random variables with* $B_{ii}\in\mathcal{N}(0,2n^{-1})$*. Then*

$$\mathbb{P}[\eta_n<t]=\left(2F\left(\sqrt{\frac{n}{2}}\,t\right)-1\right)^n, \tag{4.20}$$

where $F(x)$ *is the c.d.f. of a random variable distributed* $\mathcal{N}(0,1)$*.*

Proof. Since $\eta_n := \max_{1\leq i\leq n}\{|B_{ii}|\}$, we have

$$\mathbb{P}[\eta_n<t]=\mathbb{P}\left[\bigcap_{i=1}^{n}\{|B_{ii}|<t\}\right]=\prod_{i=1}^{n}\mathbb{P}[|B_{ii}|<t],$$

where $B_{ii}\in\mathcal{N}(0,2n^{-1})$. We therefore have

$$\mathbb{P}[|B_{ii}|<t]=\mathbb{P}\left[|\xi|<\sqrt{\frac{n}{2}}\,t\right]=\mathbb{P}\left[-\frac{\sqrt{n}t}{\sqrt{2}}<\xi<\frac{\sqrt{n}t}{\sqrt{2}}\right],$$

where $\xi\in\mathcal{N}(0,1)$. The last probability can easily be calculated numerically, since

$$\mathbb{P}\left[-\frac{\sqrt{n}t}{\sqrt{2}}<\xi<\frac{\sqrt{n}t}{\sqrt{2}}\right]=2F\left(\frac{\sqrt{n}t}{\sqrt{2}}\right)-1,$$

where $F(x)$ is the c.d.f. of ξ. Hence we obtain

$$\mathbb{P}[\eta_n<t]=\left(2F\left(\sqrt{\frac{n}{2}}\,t\right)-1\right)^n,$$

as required. □

From Proposition 4.2.1 it follows that the probability in (4.12) can be calculated using the c.d.f. of the standard normal distribution, denoted by $F(t)$. That is,

$$\mathbb{P}\left[\max_{1\leq j\leq n}|B_{jj}|\geq\frac{\min\{\gamma_{k-1},\gamma_k\}}{2\varepsilon_0}\right]=1-\left(2F\left(\sqrt{\frac{n}{2}}\,\frac{\min\{\gamma_{k-1},\gamma_k\}}{2\varepsilon_0}\right)-1\right)^n. \tag{4.21}$$

Hence, if we are given an $\varepsilon_0 > 0$ and want to bound the probability of a swap, (4.19), from above by (4.12), we have to use our own judgement of whether ε_0 is "small enough". This would determine how reliable the probability in (4.12) is as an upper bound of (4.19). In Experiment 4.4.1 we test (4.21) as an upper bound of $\mathbb{P}[\lambda_n(\varepsilon_0)$ has swapped $]$ for $n = 100$ and different values of ε_0. We also construct an example, in which (4.12) fails to be true. This shows that, whether ε_0 is small enough for this approach to be valid, also depends on n and $\min\{\gamma_{k-1}, \gamma_k\}$. While the dependance of ε_0 on $\min\{\gamma_{k-1}, \gamma_k\}$ was somewhat expected, we were surprised to discover that the accuracy of the linear approximations $\tilde{\lambda}_i(\varepsilon_0)$ of $\lambda_i(\varepsilon_0)$ also depends on the size of the matrix.

4.3. Bounding the probability of a swap from above, by combining Theorem 3.2.2 and the Bauer-Fike Theorem.

In this section we provide a solution to Problem 4.1.1 using Theorem 3.2.2 and the symmetric version of the Bauer-Fike Theorem, which we stated in §3.2. Our main result is inequality (4.25) in Theorem 4.3.1, which gives an upper bound on $\mathbb{P}[\lambda_k(\varepsilon_0)$ has swapped $]$ in terms of the cumulative distribution function of $\|B\|_2 + \|B(:,k)\|_2$, where $B(:,k)$ is the MATLAB notation for the k-th column of the matrix B. Finally, we briefly recall a way of approximating the bound in (4.25) numerically and discuss the disadvantages of the upper bound given in Theorem 4.3.1.

We start by recalling the following corollaries from §3.5.

Corollary 4.3.1. *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices, ε be a scalar and $A(\varepsilon) = A + \varepsilon B$. Let λ_k be a simple eigenvalue of A and $\mathbf{v}_k$ be its corresponding eigenvector of unit length. Then there exists an eigenvalue, $\lambda(\varepsilon)$, from the spectrum of $A(\varepsilon)$, such that*

$$|\lambda_k - \lambda(\varepsilon)| \leq |\varepsilon| \|B\mathbf{v}_k\|_2. \tag{4.22}$$

Corollary 4.3.2. *In the settings of Corollary 4.3.1, if ε is such that*

$$|\varepsilon|(\|B\mathbf{v}_k\| + \|B\|_2) < \min\{\gamma_{k-1}, \gamma_k\}, \tag{4.23}$$

then we have

$$|\lambda_k - \lambda_k(\varepsilon)| \leq |\varepsilon| \|B\mathbf{v}_k\| \qquad \text{and} \qquad |\lambda_k - \lambda_j(\varepsilon)| > |\varepsilon| \|B\mathbf{v}_k\| \quad \text{for all } j \neq k.$$

Remark 4.3.1. *Corollary 4.3.2 implies that when $\varepsilon_0 > 0$ is such that*

$$\varepsilon_0(\|B\|_2 + \|B\mathbf{v}_k\|_2) < \min\{\gamma_{k-1}, \gamma_k\},$$

then

$$\lambda_k(\varepsilon) \neq \lambda_j(\varepsilon), \quad \text{for all } \varepsilon, \text{ s.t. } |\varepsilon| \leq \varepsilon_0 \quad \text{and all } j \neq k. \tag{4.24}$$

The aim now is to extend these results to perturbation of symmetric matrices by SGOE matrices and to provide an upper bound on the probability of the event $\{\lambda_k \text{ has swapped }\}$ (see Definition 4.1.2). As it was discussed earlier, we may (without loss of generality) consider perturbations of the form

$$A(\varepsilon) := A + \varepsilon B,$$

where $A \in \mathbb{R}^{n\times n}$ is a diagonal matrix and $B \in \mathbb{R}^{n\times n}$ is SGOE matrix. Thus, one could translate Corollary 4.3.2 to the case when B is SGOE matrix in the following way.

Theorem 4.3.1. *Let $A \in \mathbb{R}^{n\times n}$ be deterministic symmetric matrix and $B \in \mathbb{R}^{n\times n}$ be SGOE matrix. Further, let λ_k be a simple eigenvalue of A and let the scalar $\varepsilon_0 > 0$ be given. Then*

$$\mathbb{P}[\lambda_k(\varepsilon_0) \text{ has swapped }] \leq \mathbb{P}[|\varepsilon_0|(\|B\|_2 + \|B(:,k)\|_2) \geq \min\{\gamma_{k-1}, \gamma_k\}], \tag{4.25}$$

where $B(:,k)$ is the MATLAB notation for the k-th column of the matrix B.

Proof. In Remark 4.3.1, (4.24) in probabilistic terms is the complement of the event $\{\lambda_k(\varepsilon_0)$ has swapped $\}$. Therefore, from Remark 4.3.1 we obtain

$$\{\lambda_k(\varepsilon_0) \text{ has swapped }\} \subset \{\varepsilon_0(\|B\|_2 + B\mathbf{v}_k) \geq \min\{\gamma_{k-1}, \gamma_k\}\},$$

which is in fact

$$\{\lambda_k(\varepsilon_0) \text{ has swapped }\} \subset \{\varepsilon_0(\|B\|_2 + B(:,k)) \geq \min\{\gamma_{k-1}, \gamma_k\}\}, \tag{4.26}$$

since the matrix A may be assumed diagonal (i.e. A may be assumed diagonal, so that its diagonal entries are its eigenvalues) and thus the unit eigenvector corresponding to λ_k, $\mathbf{v}_k$, satisfies $\mathbf{v}_k = \mathbf{e}_k$, which implies $B\mathbf{v}_k = B(:,k)$. Hence, after taking probabilities on both sides of (4.26), we finally arrive at (4.25). □

We now briefly discuss the implementation of the result of Theorem 4.3.1. In fact, most of the things we are about to say have already been mentioned in the discussion after Theorem 3.5.1 in §3.5.

Theorem 4.3.1 states that, given an $\varepsilon_0 > 0$, we can bound the probability

$$\mathbb{P}[\lambda_k(\varepsilon_0) \text{ has swapped }] \qquad \text{by} \qquad \mathbb{P}[\varepsilon_0(\|B\|_2 + \|B(:,k)\|_2) \geq \min\{\gamma_{k-1}, \gamma_k\}].$$

However, there are certain difficulties in obtaining the latter probability in practice. These difficulties are mainly due to the fact that we don't know the joint distribution of $\|B\|_2$ and $\|B(:,k)\|_2$. In the discussion following Theorem 3.5.1 we argued that the random variables $\|B\|_2$ and $\|B(:,k)\|_2$ become "less dependent" as n becomes larger. This led to approximating the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$ by assuming that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent. Numerical tests (see Experiment 3.9.2) confirmed that such an assumption gives reasonably close results to the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$ obtained by simulation. We suggest the same strategy here for the computation of the probability $\mathbb{P}[|\varepsilon_0|(\|B\|_2 + \|B(:,k)\|_2) \geq \min\{\gamma_{k-1}, \gamma_k\}]$.

If we assume that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent, then the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$, denoted by $F_{\|B\|_2+\|B(:,k)\|_2}$, can easily be calculated as a convolution of the c.d.f.'s of $\|B\|_2$ and $\|B(:,k)\|_2$ (this has been explained in greater detail in Experiment 3.9.2), where an approximation of the c.d.f. of $\|B\|_2$, via the solution to an initial value problem, was given in §2. An easy way of obtaining the c.d.f. of $\|B(:,k)\|_2$ has also been given in Experiment 3.9.2.

Therefore, as a conclusion, given an $\varepsilon_0 > 0$ we can approximate

$$\mathbb{P}[\varepsilon_0(\|B\|_2 + \|B(:,k)\|_2) \geq \min\{\gamma_{k-1}, \gamma_k\}],$$

by calculating

$$F_{\|B\|_2+\|B(:,k)\|_2}\left(\frac{\min\{\gamma_{k-1}, \gamma_k\}}{\varepsilon_0}\right),$$

under the assumption that $\|B\|_2$ and $\|B(:,k)\|_2$ are independent random variables. Then we can take $F_{\|B\|_2+\|B(:,k)\|_2}\left(\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_0}\right)$, calculated by that assumption, as an upper bound on the probability $\mathbb{P}[\lambda_k(\varepsilon_0) \text{ has swapped }]$.

Finally, we consider the advantages and disadvantages of this approach. One advantage is that the approach considered in this section does not depend on the magnitude of ε_0, the size of the perturbation. This is because the Bauer-Fike Theorem and Theorem 3.2.2 are valid for any ε. However, such a robustness comes at a certain price. The upper bound in Theorem 4.3.1 involves the c.d.f. of $\|B\|_2 + \|B(:,k)\|_2$ and thus, it should depend on the speed of convergence of

$$\|B\|_2 + \|B(:,k)\|_2 \xrightarrow{\mathcal{D}} 3, \quad \text{as } n \to \infty \tag{4.27}$$

(see §3.6). For example, suppose $\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_0}$ is kept constant and only n is increased. Then, by definition, (4.27) implies that, depending on whether $\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_0} > 3$ or $\frac{\min\{\gamma_{k-1},\gamma_k\}}{\varepsilon_0} < 3$, the upper bound given in Theorem 4.3.1 will be either one or zero, respectively. In particular, the values of $F_{\|B\|_2+\|B(:,k)\|_2}\left(\frac{\gamma_{n-1}}{\varepsilon_0}\right)$, the upper bound on $\mathbb{P}[\lambda_n(\varepsilon_0)$ has swapped $]$ from Theorem 4.3.1, for $n = 100$ and $\varepsilon_0 = 0.4$ and 0.5 (see Table 4.1 below) are all equal to one, while the values of the simulated probability of a swap are close to zero.

In its present state this analysis appears to have limited practical application. Refinements may change this perspective.

4.4. A numerical experiment.

Experiment 4.4.1. *In this experiment we compare the upper bound on*

$$\mathbb{P}[\lambda_n(\varepsilon_0) \text{ has swapped }],$$

given in §4.2 by (4.12) *with results from simulations.*

Firstly, we discuss the data which we use for this experiment. $n-1$ *of the non-zero entries of the diagonal matrix* A *are simulated as independent, uniformly distributed random variables on the interval* $(0,1)$ *and the* n*-th is chosen equal to* 1.5 *(so that* A *is* $n \times n$ *diagonal matrix). This choice of the entries of* A *ensures two things: Firstly, that the gap between the largest two eigenvalues of* A *is approximately equal to* 0.5 *and secondly, that the* 2*-norm of the matrix* A *is fixed at* 1.5*. The matrix by which we perturb* A*,* B*, is SGOE matrix. Since the* 2*-norm of* A *is fixed and* $\|B\|_2 \xrightarrow{\mathcal{D}} 2$*, as* $n \to \infty$*, the magnitude of the norms of* A *and* B *do not scale (increase) with their size. This, we think, makes first-order Perturbation Theory consistent for matrices of different sizes.*

The values of ε_0 *used for this experiment are* $\varepsilon_0 = 0.4, 0.5, \ldots, 0.9$ *(see Table 4.1 below). The value of* n *is fixed at* 100*.*

Secondly, the upper bound given in (4.12) *is calculated using formula* (4.21)*, where we recall that* $F(t)$ *denotes the probability that a random variable distributed* $\mathcal{N}(0,1)$ *is less than* t*. The result is denoted by* $\mathbb{P}_{\text{Lin}}$ *in Table 4.1.*

Thirdly, we describe how we find $\mathbb{P}[\lambda_n(\varepsilon_0)$ *has swapped* $]$ *by simulation (denoted by* $\mathbb{P}_{\text{Sim}}$*). The SGOE matrix* B *is simulated* $10\,000$ *times and for each sample we calculate* $\mathbf{v}_n(\varepsilon_0)$*, the unit eigenvector corresponding to the largest eigenvalue of* $A+\varepsilon_0 B$*,* $\lambda_n(\varepsilon_0)$*. For each sample of the vector* $\mathbf{v}_n(\varepsilon_0)$ *we compare the magnitude of its last entry,*

$|\mathbf{v}_n^{[n]}(\varepsilon_0)|$, *with* $\max_{1\leq i\leq n-1}|\mathbf{v}_n^{[i]}(\varepsilon_0)|$. *If*

$$|\mathbf{v}_n^{[n]}(\varepsilon_0)| < \max_{1\leq i\leq n-1}|\mathbf{v}_n^{[i]}(\varepsilon_0)|, \tag{4.28}$$

then we count that $\lambda_n(\varepsilon_0)$ *has swapped. Otherwise, it has not. At the end, we obtain the probability* $\mathbb{P}[\lambda_n(\varepsilon_0)$ *has swapped* $]$ *as the number of instances in which* $\lambda_n(\varepsilon_0)$ *has swapped divided by the number of simulations.*

We now explain briefly why (4.28) *is used as a criterion that* $\lambda_n(\varepsilon_0)$ *has swapped with some other eigenvalue from the spectrum of* $A(\varepsilon_0)$.

Since A *is a diagonal matrix, its eigenvalues are equal to its diagonal entries, that is,* $\lambda_i = A(i,i)$, $1 \leq i \leq n$, *where we have assumed that the diagonal entries of* A *are sorted in an ascending order. Therefore, the unit eigenvector corresponding to* λ_n, $\mathbf{v}_n$, *is equal to* $\mathbf{e}_n$ *(the vector whose entries are zeros, apart from its* n*-th entry, which is equal to one). Thus,* (4.28) *is in fact*

$$|\mathbf{v}_n^T(\varepsilon)\mathbf{v}_n| < \max_{1\leq i\leq n-1}|\mathbf{v}_n^T(\varepsilon_0)\mathbf{v}_i|,$$

where $\max_{1\leq i\leq n}|\mathbf{v}_n^T(\varepsilon)\mathbf{v}_i|$ *gives the cosine of the angle between* $\mathbf{v}_n(\varepsilon_0)$ *and the subspace spanned by the vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$. *Therefore, by* (4.28) *we compare the cosine of the angle between* $\mathbf{v}_n(\varepsilon_0)$ *and* $\mathbf{v}_n$ *with the cosine of the angle between* $\mathbf{v}_n(\varepsilon_0)$ *and the subspace orthogonal to* $\mathbf{v}_n$. *Hence, if* (4.28) *is satisfied, it is reasonable to assume that* $\mathbf{v}_n(\varepsilon_0)$ *has initially corresponded to some* $\lambda_i(\varepsilon)$ *for* $i \neq n$ *and* $\varepsilon < \varepsilon_0$. *In other words,* $\mathbf{v}_n(\varepsilon_0)$ *has initially been* $\mathbf{v}_i(\varepsilon)$, *for some* $\varepsilon < \varepsilon_0$ *and* $i \neq n$ *and thus, it has been "closer" to the subspace orthogonal to* $\mathbf{v}_n$. *Then* $\lambda_i(\varepsilon)$ *has swapped with* $\lambda_n(\varepsilon)$, *which has caused* $\mathbf{v}_i(\varepsilon)$ *to become* $\mathbf{v}_n(\varepsilon)$ *after the swap, and vice versa.*

The results in Table 4.1 show that $\mathbb{P}_{\text{Lin}}$ *is indeed an upper bounds on* $\mathbb{P}_{\text{Sim}}$ *for this value of* n.

	$\varepsilon_0 = 0.4$	$\varepsilon_0 = 0.5$	$\varepsilon_0 = 0.6$	$\varepsilon_0 = 0.7$	$\varepsilon_0 = 0.8$	$\varepsilon_0 = 0.9$
$\mathbb{P}_{\text{Sim}}$	0	1.4×10^{-3}	3.45×10^{-2}	7.35×10^{-2}	2.650×10^{-1}	4.197×10^{-1}
$\mathbb{P}_{\text{Lin}}$	6×10^{-4}	1.39×10^{-2}	2.069×10^{-1}	6.343×10^{-1}	8.861×10^{-1}	9.732×10^{-1}

Table 4.1: *Upper bounds on the probability of swap of* $\lambda_n(\varepsilon_0)$ *for* $n = 100$.

The next remark discusses the problem with the linearised theory (see §4.2).

Remark 4.4.1. *The values of* $\mathbb{P}_{\text{Lin}}$ *exceed those of* $\mathbb{P}_{\text{Sim}}$ *considerably. For example, for* $\varepsilon_0 = 0.5$ *and* 0.7 $\mathbb{P}_{\text{Lin}}$ *is almost ten times as big as* $\mathbb{P}_{\text{Sim}}$. *A disadvantage of the formula for* $\mathbb{P}_{\text{Lin}}$ *is that it relies only on first-order Perturbation Theory. Thus one has to be*

careful whether ε_0, the magnitude of the errors, is not too big for the linear expansions to be valid. Another disadvantage is that the rate of convergence of $\max_{1\leq i\leq n}|B_{ii}|$ to zero is, "loosely speaking", like $n^{-1/2}$. Therefore, from the formula

$$\mathbb{P}_{\text{Lin}} = \mathbb{P}\left[\max_{1\leq i\leq n}|B_{ii}| \geq \frac{\gamma_{n-1}}{2\varepsilon_0}\right], \tag{4.29}$$

it follows that the values of $\frac{\gamma_{n-1}}{2\varepsilon_0}$ have to be close to $n^{-1/2}$, in order for this upper bound to produce meaningful results. Hence, if the gap γ_{n-1} is fixed, this forces ε_0 to grow with n, which contradicts the fact that it has to be "small" for the linearisations of $\lambda_n(\varepsilon_0)$ to be reliable. Also, if the ratio $\frac{\gamma_{n-1}}{\varepsilon_0}$ is kept fixed and only n is increased, intuitively, the probability of a swap of $\lambda_n(\varepsilon_0)$ should increase, since the eigenvalues of the the matrix A should become denser (if A's norm is preserved) and thus, the gaps γ_i, $1 \leq i \leq n-2$ will decrease. On the other hand, the probability $\mathbb{P}_{\text{Lin}}$, given in (4.29), will decrease, since

$$\max_{1\leq i\leq n}|B_{ii}| \xrightarrow{\mathcal{D}} 0.$$

For example, complementary to the results in Table 4.1, we tested the values of $\mathbb{P}_{\text{Lin}}$ and $\mathbb{P}_{\text{Sim}}$ for $n = 500$ and $\varepsilon_0 = 0.9$. In this case we obtained

$$\mathbb{P}_{\text{Lin}} = 0.0046 \qquad \textit{and} \qquad \mathbb{P}_{\text{Sim}} = 0.3321,$$

which means that $\mathbb{P}_{\text{Lin}}$ is not an upper bound on $\mathbb{P}_{\text{Sim}}$ in this case.

Chapter 5. Analytical and numerical results for entrainment in large networks of coupled oscillators.

5.1. Introduction.

In this chapter we apply knowledge of networks and related matrix theory to the problem of analysing entrainment in networks of coupled oscillators. The seminal paper in that topic is that of Kuramoto (Kuramoto, 1975), who considered networks of oscillators, in which the coupling between every pair of oscillators was identical. Although simple at a glance, his model was hard to analyse but due to his ingenious heuristics and assumptions, he was able to derive some properties about the system he considered. We consider a more general class of range dependent networks where the pairwise coupling is a probabilistic function of distance (range) between the nodes (c.f. (Grindrod, 2002)), and each node represents an oscillator with its own intrinsic phase and natural frequency of oscillation. Range dependent networks exhibit the "small world" phenomenon, being effectively superpositions of many networks each operating over different range lengths. We provide an asymptotic analysis in terms of a network coupling parameter that gives a simple analytic description of entrainment in networks of coupled oscillators. Numerical experiments are presented that agree well with the theory. The analysis is then applied to the case of a coupled system of oscillators exhibiting a "master-slave" relationship. Again numerical results agree well with the theory.

5.2. Oscillators coupled via a directed graph.

By entrainment (or synchronisation) of a system of oscillators, we mean a state of the system, in which all oscillators move together as one with a possible difference in their phases, which remains constant for large time. This is a key concept in the understanding of self-organisation phenomena of coupled oscillators (see, for example,

(Kuramoto, 1984)).

We next discuss how the Perron-Frobenius matrix theory may be used to uncover a "master-slave" structure in a network and then how this knowledge can be utilised in the entrainment analysis. Numerical experiments again support the validity of the analysis.

First, consider the simplest case of two coupled oscillators:

$$\dot{\theta}_1 = \lambda_1 + \varepsilon A_{12} \sin(\theta_2 - \theta_1)$$
$$\dot{\theta}_2 = \lambda_2 + \varepsilon A_{21} \sin(\theta_1 - \theta_2)$$

which has state space the torus with coordinates $\theta_i \mathrm{mod}(2\pi)$ for $i = 1, 2$. Here the A_{12} and A_{21} are nonnegative coupling coefficients; ε is a nonnegative overall "strength" parameter to scale the coupling; and the λ_1, λ_2 represent the uncoupled positive frequencies of the separate oscillators. Setting $\phi = \theta_2 - \theta_1$ we obtain a single equation for the phase difference:

$$\dot{\phi} = \lambda_2 - \lambda_1 - \varepsilon(A_{12} + A_{21}) \sin \phi, \tag{5.1}$$

which is integrable and so a closed form solution is available. However, qualitative information about the oscillation can be obtained directly from (5.1). First note that the frequencies become entrained for large time (with ϕ tending to a stable rest point) if and only if ε is such that

$$|\lambda_2 - \lambda_1| < \varepsilon(A_{12} + A_{21}).$$

If this condition does not hold one of the oscillators repeatedly "laps" the other.

Let us generalise the above situation to N coupled oscillators. We shall think of them as vertices connected by a directed graph with entraining couplings defining the non negative weights of directed edges. Each oscillator is represented by a single phase variable, $\theta_i \mathrm{mod}(2\pi)$, having a natural, uncoupled frequency, whilst each coupling term, say from oscillator k acting on oscillator i, affects to increase or retard the rate of increase of the phase of oscillator i, so as to approach the phase of oscillator k. The state space for the full coupled system is an N dimensional torus with coordinates $\theta_i \mathrm{mod}(2\pi)$ for $i = 1, \ldots, N$. Specifically, we consider the following system on the N-torus:

$$\dot{\theta}_i = \lambda_i + \varepsilon \sum_{k=1}^{N} A_{ik} \sin(\theta_k - \theta_i), \quad i = 1, \ldots, N. \tag{5.2}$$

Introduce the $n \times n$ coupling matrix A with zeros on the diagonal and (i, k)-th component, A_{ik}, which represents the weight of the coupling, or edge, from vertex k to

vertex i. The parameter ε is a nonnegative overall "strength" parameter to scale the impact of A; and the λ_i represent the positive uncoupled frequencies of the separate oscillators.

Our interest is in whether and how the oscillators can become entrained with one another, for large time; producing a baulk oscillation, namely their phases moving together, possibly separated by a constant set of phase shifts. Like the simple twin-oscillator case, this behaviour depends upon the strength and nature of the couplings as well as the distribution of their natural frequencies.

5.2.1. No Baulk Oscillations for small ε.

For any i and j we have

$$\dot{\theta}_i - \dot{\theta}_j = \lambda_i - \lambda_j +$$

$$\varepsilon \left(\sum_{k=1}^{N} A_{ik} \sin(\theta_k - \theta_i) - \sum_{k=1}^{N} A_{jk} \sin(\theta_k - \theta_j) \right).$$

The left hand side of this equation must vanish when oscillators i and j are entrained (that is when their phases differ by a constant amount through time). Set

$$\varepsilon^* = \max_{1 \le i,j \le N} \frac{|\lambda_i - \lambda_j|}{\sum_{k=1}^{N} (A_{ik} + A_{jk})}. \tag{5.3}$$

Then if $\varepsilon < \varepsilon^*$, $\dot{\theta}_i = \dot{\theta}_j$ is impossible for at least one pair of oscillators and there can be no baulk oscillation. Note this condition is necessary and sufficient for no baulk oscillation to exist when $N = 2$.

5.2.2. Asymptotic Analysis of Baulk Oscillations for large ε.

We seek an asymptotic solution, valid in the limit of large ε, representing a baulk oscillation, so that for some function, $\theta_0(t)$ say, we have

$$\theta_i(t) = \theta_0(t) + \text{ an } \varepsilon\text{-dependent phase shift for oscillator } i$$

for each $i = 1, \ldots, N$.

Setting $\boldsymbol{\theta}(t) = [\theta_1(t), \theta_2(t), \ldots, \theta_N(t)]^T$, $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$ and $\mathbf{1} = [1, 1, \ldots, 1]^T \in \mathbb{R}^N$, we shall seek a solution which is in the form of a baulk oscillation (that is, all phases entrained) where the phase shifts are represented by a regular expansion in

inverse powers of ε:

$$\boldsymbol{\theta}(t) = \theta_0(t)\mathbf{1} + \frac{1}{\varepsilon}\boldsymbol{\theta}_1 + \frac{1}{\varepsilon^2}\boldsymbol{\theta}_2 + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right). \tag{5.4}$$

Here $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vectors orthogonal to $\mathbf{1}$, so that the individual phase shifts are distinct from the baulk oscillation term.

Substituting (5.4) into (5.2) and expanding out the sine terms, we obtain

$$\dot{\theta}_0\mathbf{1} = \boldsymbol{\lambda} + \triangle\boldsymbol{\theta}_1 + \frac{1}{\varepsilon}\triangle\boldsymbol{\theta}_2 + \mathcal{O}\left(\frac{1}{\varepsilon^2}\right). \tag{5.5}$$

Here $\triangle$ is such that $-\triangle$ is the Laplacian matrix (see §1) associated with the network coupling matrix A (replacing the zeroes on the diagonal of A with the negative of the corresponding row sums):

$$\triangle = A - \mathrm{diag}(A\mathbf{1}).$$

The matrix $\triangle$, just like the Laplacian matrix of the network, contains information about the connected nature of the network: it is of huge importance in graph theory (Bollobas, 1995). It is easy to see that zero is an eigenvalue of $\triangle$ with multiplicity equal to the number of distinct connected sub networks. Without loss of generality we shall assume zero is a simple eigenvalue - otherwise we may consider each connected sub network separately. In that case $\triangle\mathbf{1} = \mathbf{0}$.

Let $\mathbf{e}$ denote the corresponding left unit eigenvector: $\mathbf{e}^T\triangle = \mathbf{0}^T$. Then pre-multiplying (5.5) with $\mathbf{e}^T$ we have

$$\dot{\theta}_0\mathbf{e}^T\mathbf{1} = \mathbf{e}^T\boldsymbol{\lambda} + \mathcal{O}\left(\frac{1}{\varepsilon^2}\right), \tag{5.6}$$

which determines $\theta_0(t)$. (In the case when $\triangle$ is a symmetric matrix the term $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ in the right hand side of (5.6) vanishes.) Then to $\mathcal{O}(1)$ and $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ we have $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively, determined as the unique solutions, orthogonal to $\mathbf{1}$, of the matrix equations:

$$\left(\frac{\mathbf{e}^T\boldsymbol{\lambda}}{\mathbf{e}^T\mathbf{1}}\right)\mathbf{1} - \boldsymbol{\lambda} = \triangle\boldsymbol{\theta}_1, \qquad \triangle\boldsymbol{\theta}_2 = \mathbf{0}. \tag{5.7}$$

First note that $\boldsymbol{\theta}_2 = \mathbf{0}$. Next, we may write

$$\boldsymbol{\theta}(t) = \left(t\left(\frac{\mathbf{e}^T\boldsymbol{\lambda}}{\mathbf{e}^T\mathbf{1}}\right) + C\right)\mathbf{1} + \frac{1}{\varepsilon}\boldsymbol{\theta}_1 + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right), \tag{5.8}$$

where C is a constant, and $\boldsymbol{\theta}_1$ can be found by solving (5.7) in the subspace orthogonal to $\mathbf{1}$. For later use we call $\boldsymbol{\theta}_1$ the *first order entrainment vector*.

Hence by calculating $\mathbf{e}^T$, the left eigenvector of $\triangle$ and solving for $\boldsymbol{\theta}_1$ from (5.7), we can use (5.8) to estimate the behaviour of the oscillators for large coupling parameter ε. In fact the experiments in the next section show that ε does not need to be too large. Indeed, for values of ε not too much greater than ε^*, (5.8) provides an accurate representation of the behaviour of the system.

Finally, we note that for the network example considered in the next section, the second eigenvalue of $\triangle$ is small (equalling -0.01528) with corresponding eigenvector $\mathbf{v}_F$, often called the Fiedler vector (Fiedler, 1975). Hence $\boldsymbol{\theta}_1$ will typically be rich in the direction of $\mathbf{v}_F$. Now $\mathbf{v}_F$ is often used to explain certain network features (for example, clustering) and this suggests that the Fiedler vector might also provide information to help understand different features in the solutions of (5.2).

5.3. Numerical Example: entrainment for range dependant coupling.

We take $N = 100$, A a symmetric random range dependent matrix with values lying between zero and 0.96, and the λ_i as independent uniformly distributed random numbers within the interval $[0.5; 1.5]$. Then by direct calculation, $\varepsilon^* = 0.47884$.

In this case $\mathbf{e} = \frac{1}{\sqrt{N}}\mathbf{1}$ and so we have from (5.6)

$$\dot{\theta}_0 = \frac{1}{N}\sum_{i=1}^{N} \lambda_i =: \hat{\lambda}.$$

Hence (5.8) gives

$$\theta_i(t) = \hat{\lambda}t + C + \frac{1}{\varepsilon}\theta_1^{[i]} + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right),$$

where $\theta_1^{[i]}$ denotes the i-th component of $\boldsymbol{\theta}_1$, and

$$\theta_i(t) - \theta_j(t) = \frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[j]}) + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right). \tag{5.9}$$

In Figure 5-1 we plot the phase differences, $\theta_i(t) - \theta_1(t)$ for $i = 2, \ldots, 100$, obtained directly from the numerical solution, for $t \in [0; 50]$, for various values of ε ($\varepsilon = 0.5, 0.6, 0.8, 2.0, 5.0, 10.0$).

The entrainment as ε increases is clearly seen in Figure 5-1. Indeed, for $\varepsilon = 2.0$ the system settles to baulk oscillation before $t = 250$. In Figure 5-2 we compare the values of $\theta_i(t) - \theta_1(t)$ obtained by numerical solution with $\frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[1]})$ in order to test the validity of (5.9), and hence the validity of the asymptotic analysis leading to

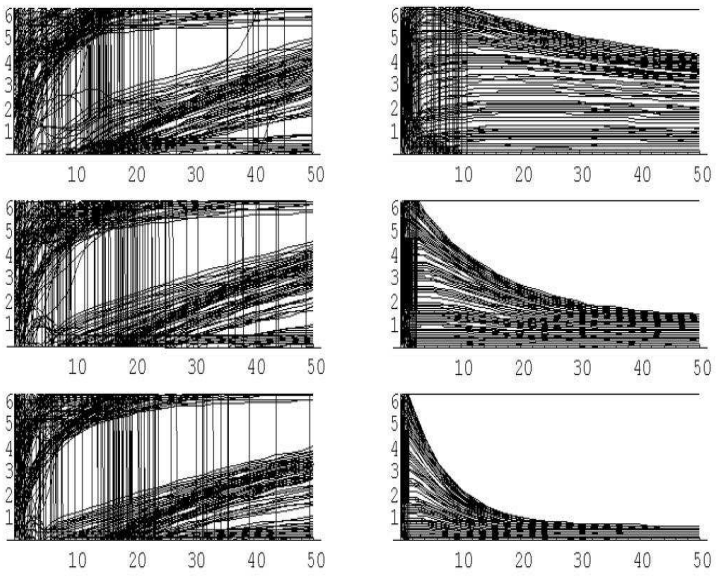

Figure 5-1: *Plot of* $\theta_i - \theta_1$, *for* $i = 2, \ldots, 100$, *versus time* t, *for* $\varepsilon = 0.5, 0.6, 0.8, 2.0, 5.0, 10.0$.

equation (5.8). Clearly, even for ε not so large there is very good agreement between the asymptotic expression and numerical experiment, with the maximum error being around 1.3×10^{-4}.

Lastly, in Figure 5-3 we show the solution behaviour for the system with $\varepsilon = \varepsilon^* + 0.03$ for random starting values. This Figure represents the plot of the terms $\theta_i(t) - \theta_{i_0}(t)$, where $t \in [0; 250]$ and i_0 is such that $\lambda_{i_0} \leq \lambda_i$ for $1 \leq i \leq 100$. In our simulation in Figure 5-3 we observe that there are two clusters of oscillators entrained with θ_{i_0} and two other clusters which drift away from them. There is an "extreme" oscillator, which is not entrained to any of the groups, and two other oscillators, which seem to be attracted by the clusters of oscillators.

5.4. A "master-slave" system.

In this section we consider a special case of the coupled system (5.2), namely, where the coupling is such that there are m "master" and s "slave" oscillators, with directed edges from the master set to the slave set, but no directed edges from the slaves back to the masters. We describe this situation as follows: Assume (5.2) can be written in

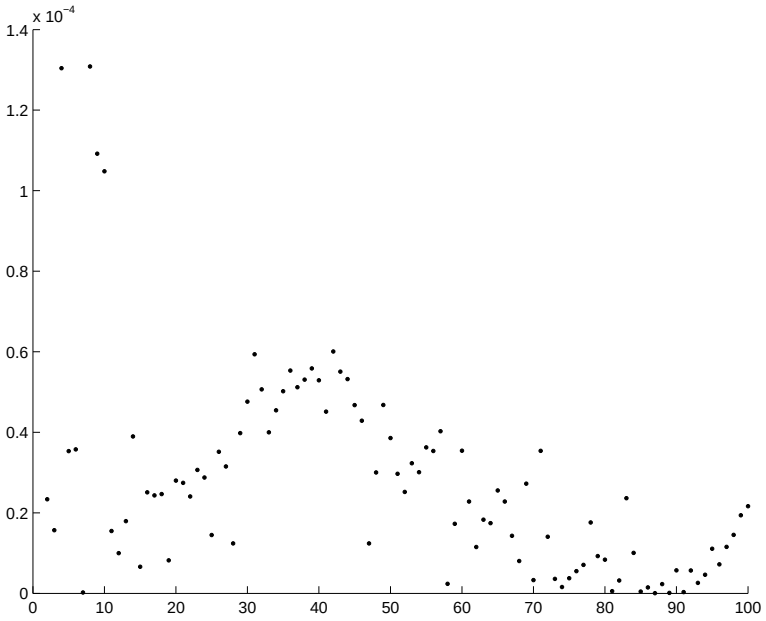

Figure 5-2: *In this Figure we plot, for* $i = 2, \ldots, 100$*, the absolute value of the difference between* $\theta_i(t) - \theta_1(t)$ *(obtained by numerical solution of* (5.2)*) and* $\frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[1]})$*, see* (5.9)*. Here* $\varepsilon = 2$ *and* $t = 250$*.*

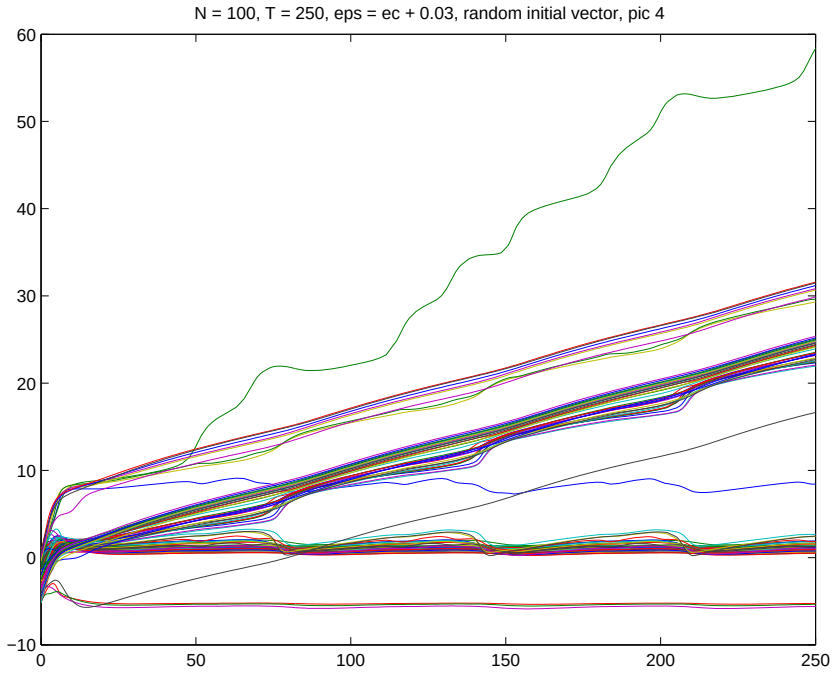

Figure 5-3: *Plot of* $\theta_i - \theta_{i_0}$ *versus time, for* $\varepsilon = \varepsilon^* + 0.03$*.*

the form

$$\dot{\theta}_i = \lambda_{m,i} + \varepsilon \sum_{k=1}^{m} B_{ik} \sin(\theta_k - \theta_i) \tag{5.10}$$

for $i = 1, 2, \ldots, m$, and

$$\dot{\psi}_j = \lambda_{s,j} + \varepsilon \sum_{k=1}^{m} C_{jk} \sin(\theta_j - \psi_k) + \varepsilon \sum_{k=1}^{s} D_{jl} \sin(\psi_l - \psi_j) \tag{5.11}$$

for $j = 1, 2, \ldots, s$. Here B is an $m \times m$ coupling matrix with zeros everywhere except that the (i,k)-th component equal to 1 if there is an edge from k to i; C is an $s \times m$ coupling matrix with zeros everywhere except that the (j,k)-th component equal to 1 if there is an edge from k to j, and D is an $s \times s$ coupling matrix with zeros everywhere except that the (j,l)-th component is equal to 1 if there is an edge from l to j. Hence the oscillators $\theta_1, \theta_2, \ldots, \theta_m$ in (5.10) form a self-contained subsystem, but, in addition, affect the oscillators $\psi_1, \psi_2, \ldots, \psi_s$ in (5.11) through the coupling matrix C. The oscillators $\theta_1, \theta_2, \ldots, \theta_m$ are called the "masters"; $\psi_1, \psi_2, \ldots, \psi_s$ are called the "slaves". The quantities $\lambda_{m,i}$, $1 \leq i \leq m$ and $\lambda_{s,j}$, $1 \leq j \leq s$ represent the uncoupled frequencies of the master and slave oscillators, respectively.

We shall discuss two questions in this section. First, given a general system of coupled oscillators, how do we recognise that the system has a "master-slave" structure and consequently split the system into the form given by (5.10), (5.11)? Second, can we make use of the "master-slave" structure in the system to split and possibly simplify the entrainment analysis given in §5.2? We answer these questions in the following two subsections, and in §5.5 we provide a numerical example.

5.4.1. Detecting the "master-slave" structure.

The coupling matrices in (5.2), (5.10) and (5.11) have positive or zero elements. Such matrices are said to be non-negative and are denoted $A \geq 0$, etc. If all elements of A are positive, we write $A > 0$. There is a rich theory for such matrices, see for example (Gantmacher, 1959). In particular, a square matrix is *irreducible* if there is no permutation matrix P such that

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square matrices. A matrix is *reducible* if it is not irreducible. One way to determine if a "master-slave" relationship in (5.2) exists is to determine if there exists a permutation matrix P such that the matrix A can be reduced to the

form

$$P^T A P = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix}, \tag{5.12}$$

where B and D are square matrices. Such a form, provided there exists at least one pair (i, j) such that $C(i, j) \neq 0$, would mean that there are links from some of the elements in B to elements in D. On the other hand, (5.12) would also mean that there are no links from the elements of D to the elements of B, since the top right block-matrix has only zero elements. In other words, we could solve the dynamics of the system determined by the matrix B separately and this would then determine the dynamics of the rest of the system. This can be seen from (5.10) and (5.11), where the first system of equations, (5.10), represents the dynamics of the "master" oscillators and is independent of the second system, (5.11), which gives the dynamics of the "slave" oscillators.

Of course, B and D may themselves be reduced further, but in this chapter, for simplicity, we shall assume that A is reducible and that B is an $m \times m$ irreducible matrix and D is an $s \times s$ irreducible matrix. We shall also assume that the spectral radius of A, denoted r, is simple, and shall denote the corresponding left and right eigenvectors by $\mathbf{u}$ and $\mathbf{v}$, that is,

$$\mathbf{u}^T A = r\mathbf{u}^T \quad \text{and} \quad A\mathbf{v} = r\mathbf{v}.$$

A fundamental result for non-negative matrices is the following Theorem.

Theorem 5.4.1 (Perron-Frobenius)**.** *(c.f. (Gantmacher, 1959), p. 79) Assume the spectral radius, r, of a matrix $A \geq 0$ is simple.*

(a) *If $\mathbf{u} > 0$ and $\mathbf{v} > 0$, then A is irreducible.*

(b) *Conversely, if $\mathbf{u}$ or $\mathbf{v}$ has zero components, then A is reducible.*

It is part (b) of this theorem that suggests a way of determining the reducible structure of a non-negative matrix, namely, compute $\mathbf{u}$ and $\mathbf{v}$ and match their positive components. Thus for a matrix A, which could be permuted to the form (5.12) with B and D irreducible, we would find precisely m components of $\mathbf{u}$ and $\mathbf{v}$ with $u_i \neq 0$, $v_i \neq 0$. The indices of these m components will determine the permutation matrix P that would permute A to the form (5.12).

Remark 5.4.1. *For a general matrix $A \geq 0$ one must first remove any rows and columns containing all zeros, since these represent "slaves" and "masters" that have a rather simple connectivity to the network. (This step may need to be done recursively.)*

Next, one should use part (b) of the Perron-Frobenius Theorem. If B and D are also reducible, then one needs to repeat the process.

To show the applicability of this technique, we test it on a simple adjacency matrix derived from the network given in Figure 5-4.

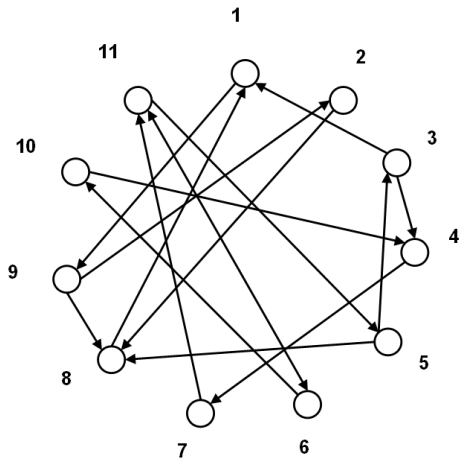

Figure 5-4: *Here a directed edge from one vertex to another is denoted by $\rightarrow$, or $\leftrightarrow$ if there is an edge both ways.*

This network produces an adjacency matrix of the form

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.13}$$

that has spectral radius $r = 1.3479$, with $\mathbf{u}$ and $\mathbf{v}$ (from the MATLAB function eigs) given by

$$\mathbf{u} = [0.5, 0.3, 0.1, 0.2, 0.1, 0.1, 0.1, 0.6, 0.4, 0.1, 0.2]^T$$

and

$$\mathbf{v} = [0, 0, 0.2, 0.3, 0.2, 0.6, 0.4, 0, 0, 0.2, 0.5]^T.$$

Matching nonzero components of $\mathbf{u}$ and $\mathbf{v}$ we see that components $3, 4, 5, 6, 7, 10, 11$ are nonzero in both vectors. Hence in this case a possible permutation P such that $P^T AP$ has the form (5.12) is one which swaps rows (columns) 1 and 2 with rows (columns) 10 and 11. For this example the B and D matrices are irreducible.

We note that (Nagaraj et al., 2004) used Perron-Frobenius theory to help analyse the strength of the connectivity of nodes in a network.

5.4.2. Entrainment in the "master-slave" system (5.10), (5.11).

Once the "master-slave" structure given by (5.10), (5.11) has been discovered, it is natural to look for a similar "master-slave" relationship in the entrainment analysis of §5.2. First, let us consider the master system (5.10) only. The analysis of §5.2 applies with $\triangle_m = B - \mathrm{diag}(B\mathbf{1}_m)$, where $\mathbf{1}_m$ is the m-dimensional vector with each component being equal to 1. With $\mathbf{e}_m^T \triangle_m = \mathbf{0}^T$, and λ_m denoting the uncoupled frequencies of the master oscillators, we have, following (5.7),

$$\triangle_m \boldsymbol{\theta}_{m,1} = \left(\frac{\mathbf{e}_m^T \boldsymbol{\lambda}_m}{\mathbf{e}_m^T \mathbf{1}_m} \right) \mathbf{1}_m - \boldsymbol{\lambda}_m, \tag{5.14}$$

with $\boldsymbol{\theta}_{m,1}$ being the first order entrainment vector, i.e. the vector in the ε^{-1}-term in expansion (5.8).

For the system (5.10), (5.11) we write the masters as $\boldsymbol{\theta}_m = (\theta_1, \theta_2, \ldots, \theta_m)^T$, the slaves as $\boldsymbol{\psi}_s = (\psi_1, \psi_2, \ldots, \psi_s)^T$ and write the first order entrainment vector as

$$(\boldsymbol{\theta}_{m,1}^T, \boldsymbol{\psi}_{s,1}^T). \tag{5.15}$$

Now the Laplacian matrix corresponding to the "master-slave" system (5.10), (5.11) can be written as

$$\triangle^{(ms)} = \begin{bmatrix} \triangle_m & 0 \\ C & \triangle_s \end{bmatrix},$$

where $\triangle_m$ is defined above and

$$\triangle_s = D - \mathrm{diag}(C\mathbf{1}_m + D\mathbf{1}_s).$$

Note that, generically, $\triangle_s$ will be nonsingular because of the addition of $\mathrm{diag}(C\mathbf{1}_m)$ to the singular matrix $D - \mathrm{diag}(D\mathbf{1}_s)$. Also $\triangle^{(ms)}\mathbf{1}_{m+s} = \mathbf{0}_{m+s}$ and $(\mathbf{e}_m^T, \mathbf{0}^T)$ is the corresponding left eigenvector. Now, if (5.7) is written in the "master-slave" notation

we obtain

$$\triangle^{(ms)} \begin{bmatrix} \boldsymbol{\theta}_{m,1} \\ \boldsymbol{\psi}_{s,1} \end{bmatrix} = \left(\frac{\mathbf{e}_m^T \boldsymbol{\lambda}_m}{\mathbf{e}_m^T \mathbf{1}_m} \right) \mathbf{1}_{m+s} - \begin{bmatrix} \boldsymbol{\lambda}_m \\ \boldsymbol{\lambda}_s \end{bmatrix}, \tag{5.16}$$

and the structure of $\triangle^{(ms)}$ means that we can decompose the calculation of (5.15) into two separate smaller calculations. First, it is straightforward to show that $\boldsymbol{\theta}_{m,1}$ in (5.16) is precisely the same vector as in (5.14), that is, the first m components of the first order entrainment vector for the system (5.10), (5.11) are precisely those given by first order entrainment vector of the "master" subsystem. Next, once $\boldsymbol{\theta}_{m,1}$ is known, the calculation of $\boldsymbol{\psi}_{s,1}$ reduces to the solution of

$$\triangle_s \boldsymbol{\psi}_{s,1} = -C\boldsymbol{\theta}_{m,1} + \left(\frac{\mathbf{e}_m^T \boldsymbol{\lambda}_m}{\mathbf{e}_m^T \mathbf{1}_m} \right) \mathbf{1}_s - \boldsymbol{\lambda}_s. \tag{5.17}$$

We note that $\boldsymbol{\psi}_{s,1}$ depends on $\boldsymbol{\lambda}_s$, $\boldsymbol{\lambda}_m$ and $\boldsymbol{\theta}_{m,1}$, and so $\boldsymbol{\psi}_{s,1}$ is itself a slave to $\boldsymbol{\theta}_{m,1}$. Thus we see that entrainment properties in the master system feed directly into the entrainment properties of the slaves. In large systems this feature may help to reduce computational costs, by first solving for $\boldsymbol{\theta}_{m,1}$ and then $\boldsymbol{\psi}_{s,1}$, rather than solving the full system.

Finally we remark that the entrainment analysis holds in the limit of large ε. As ε is reduced there are several possibilities. For example, if entrainment is lost in the master oscillators, one might expect the slaves to lose entrainment also. On the other hand, if entrainment were lost in the slave oscillators one would not necessarily expect the master set to lose entrainment. However, the strength of the coupling matrix C will play an important role in any specific application.

5.5. Numerical Example: a simple "master-slave" system.

In order to check the analysis in §5.4.2 we consider a small system of coupled oscillators of the form (5.2), where the coupling matrix A is given by (5.13), that is, $N = 11$. As described above, the matrix A can be permuted to the form (5.12). Hence a "master-slave" system of the form (5.10), (5.11) is obtained with 7 masters and 4 slaves.

For our numerical experiments we took λ_i, $1 \leq i \leq 11$ as random variables uniformly distributed in the interval $[0.5, 1.5]$. The initial conditions of the system are all set to zero, since the asymptotic theory developed in the previous sections does not depend on the initial conditions of the system. For ease of explanation of the numerical results we return to the notation of §5.2 and §5.3, with the masters being denoted by $\theta_1, \ldots, \theta_7$ and the slaves by $\theta_8, \ldots, \theta_{11}$. In Figure 5-5 we compare the phase-differences $\theta_i - \theta_1$, $2 \leq i \leq 11$ from the asymptotic analysis and the corresponding phase-differences

obtained from the MATLAB ODE solver. The entries in Figure 5-5 represent the absolute value of the difference between the numerical solution and the solution using the asymptotic theory, with • denoting the differences between $\theta_i - \theta_1$, $i = 2, \ldots, 7$ and ∘ denoting the differences between $\theta_i - \theta_1$, $i = 8, \ldots, 11$.

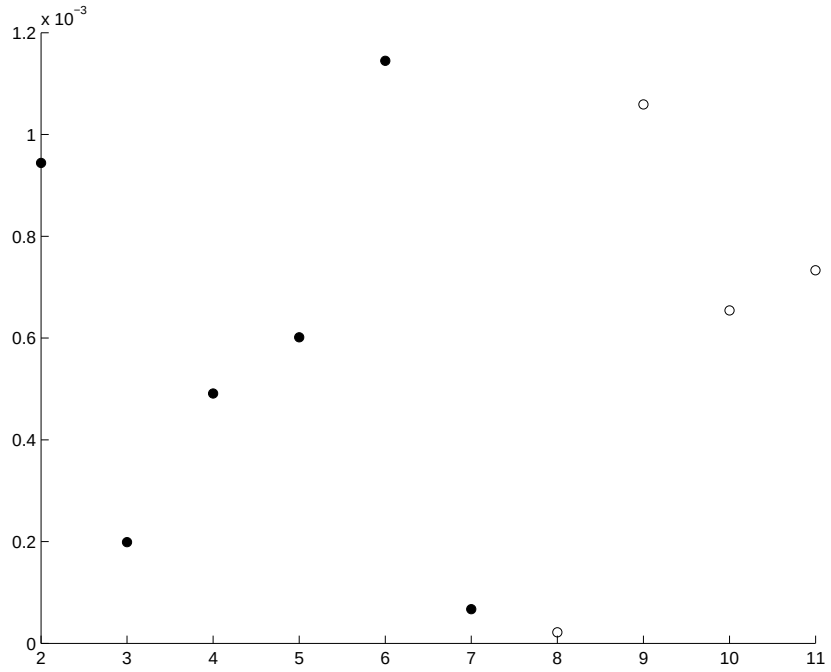

Figure 5-5: *Plot of the absolute value of the difference between numerical and asymptotic solution for* $\theta_i - \theta_1$, $2 \leq i \leq 11$. *Here* $\varepsilon = \varepsilon^* + 2$, *where* $\varepsilon^* = 0.33728$ *and is taken over the whole network.*

We see that the differences between the numerical and analytical experiments are very small and support the theory, even though ε is rather small.

Chapter 6. Clustering products of Path graphs with respect to different Laplacian matrices

6.1. Introduction

In this chapter we recall the definitions of Laplacian and normalised Laplacian matrices of graphs. We also mention a third matrix, associated with graphs, which is related to the Laplacian and normalised Laplacian matrices. We start with a brief review of spectral clustering of graphs with respect to the second eigenvector, also known as the *Fiedler vector*, of their Laplacian and normalised Laplacian matrices. Then, in order to compare the clusterings with respect to these two matrices, we introduce a Homotopy between them. In particular, we consider the Homotopy between the Laplacian and the normalised Laplacian matrices of products of Path graphs. It turns out, that even in that case, it is difficult to obtain a definitive conclusion from the comparison between these two methods of clustering. However, as our main result (in §6.4) we describe the possible ways of clustering of products of Path graphs with respect to either their Laplacian, or normalised Laplacian matrix.

Finally, we make a remark about the notation used in this chapter. Firstly, graphs are generally denoted by G or H, but we have also used the notation P_m or P_n to mean Path graphs of order m or n, respectively. Secondly, the (i, j)-th entry of the matrix L, for example, is sometimes denoted as L_{ij}, and when this notation was considered rather clumsy or confusing, we have used $L(i, j)$ instead. The same strategy has been used in the notation of the entries of vectors, although in the proof of Corollary 6.3.2 we also use $\mathbf{z}^{[j]}$ to mean the j-th entry of the vector $\mathbf{z}$. Finally, in order to emphasize the relation between a given matrix, vector, or a set of objects, and the graph with which it is associated, we have used subscripts. For example, the Laplacian matrix of the graph G is denoted by L_G and the set of vertices of the graph G is denoted by V_G.

Let $G = (V_G, E_G)$ be a graph with $V_G = \{v_1, v_2, \ldots, v_n\}$ being its set of vertices and E_G its set of edges. In §1 we introduced the Laplacian matrix of G (c.f. Defini-

tion 1.2.12), which we shall denote here by L_G. We recall that L_G's entries are given by

$$L_G(i,j) = \begin{cases} -w(v_i, v_j) & \text{if } i \neq j; \\ \sum_{k=1, k\neq i}^{n} w(v_i, v_k) & \text{if } i = j, \end{cases} \tag{6.1}$$

where $w(v_i, v_j) \geq 0$ is the *weight* of the edge joining vertices v_i and v_j, $1 \leq i < j \leq n$ (see Definition 1.2.6, where we define the term *weight function*).

Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of $L_G \in \mathbb{R}^{n\times n}$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be their corresponding eigenvectors of unit length. Here we mention some properties of the Laplacian matrix L_G, which we shall use later in this chapter. Some of these properties were proved in §1.

Properties 6.1.1. (Laplacian matrix)

1. For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ we have

$$\mathbf{x}^T L_G \mathbf{x} = -\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} L_G(i,j)(x_i - x_j)^2$$

   (c.f. Proposition 1.2.2);

2. The matrix L_G is positive semi-definite (this follows from Property 1);

3. The eigenvalues of L_G satisfy

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \qquad \text{and} \qquad \mathbf{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1},$$

   where $\mathbf{1} \in \mathbb{R}^n$ is the vector whose entries are all equal to one (this can be derived easily from Property 1);

4. The eigenvalue $\lambda_2 > 0$ if and only if G is connected;

5. The eigenvector corresponding to λ_2, $\mathbf{v}_2 = (\mathbf{v}_2(1), \mathbf{v}_2(2), \ldots, \mathbf{v}_2(n))^T$, is called *Fiedler vector* (c.f. Definition 1.2.13). By splitting the vertices of the graph into two disjoint sets, $V_G^{(1)}$ and $V_G^{(2)}$, in the following way:

$$V_G^{(1)} := \{v_i \mid \mathbf{v}_2(i) \leq 0\} \qquad \text{and} \qquad V_G^{(2)} := V_G \setminus V_G^{(1)}, \tag{6.2}$$

   we obtain a clustering of G, for which, loosely speaking, the quantity

$$\sum_{v_i \in V_G^{(1)}, v_j \in V_G^{(2)}} w(v_i, v_j)$$

is small and the quantities

$$\sum_{v_i,v_j\in V_G^{(1)}} w(v_i,v_j) \qquad \text{and} \qquad \sum_{v_i,v_j\in V_G^{(2)}} w(v_i,v_j)$$

are large (c.f. (Fiedler, 1975)).

Now we shall introduce two other matrices, also related to the graph G. Both of them are known to satisfy property similar to Property 5 of the Laplacian matrix. Our task in this chapter will be to compare the clusterings produced by these three matrices associated with G. We shall do that on Products of Path graphs, which shall be defined in §6.3.

Given a graph G, let the diagonal matrix D_G be defined by

$$D_G(i,i) := \sum_{k=1,k\neq i}^{n} w(v_i,v_k), \quad 1 \leq i \leq n.$$

In other words, $D_G(i,i) = L_G(i,i)$, for all $1 \leq i \leq n$. We recall (c.f. Definition 1.2.14) that the matrix $\hat{L}_G$ given by

$$\hat{L}_G := D_G^{-\frac{1}{2}} L_G D_G^{-\frac{1}{2}} \tag{6.3}$$

is called *normalised Laplacian matrix* of G.

Let $\hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \cdots \leq \hat{\lambda}_n$ be the eigenvalues of $\hat{L}_G$ and $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_n$ be their corresponding unit eigenvectors. Then the normalised Laplacian matrix $\hat{L}_G$, similarly to the Laplacian matrix L_G, satisfies the following properties (c.f. (Chung, 1997)):

Properties 6.1.2. (Normalised Laplacian matrix)

1. For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ we have

$$\mathbf{x}^T \hat{L}_G \mathbf{x} = -\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} L_G(i,j) \left(\frac{x_i}{\sqrt{D_G(i,i)}} - \frac{x_j}{\sqrt{D_G(j,j)}} \right)^2,$$

   where $L_G(i,j)$ is the (i,j)-th entry of the Laplacian matrix, L.

2. The matrix $\hat{L}_G$ is positive semi-definite;

3. The eigenvalues of $\hat{L}_G$ satisfy

$$0 = \hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \cdots \hat{\lambda}_n \leq 2 \qquad \text{and} \qquad \hat{\mathbf{v}}_1 = \frac{1}{\sqrt{\sum_{i=1}^{n} D_G(i,i)}} D_G^{\frac{1}{2}} \mathbf{1};$$

4. The eigenvalue $\hat{\lambda}_2 > 0$ if and only if G is connected;

5. The vector $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2 = D_G^{-\frac{1}{2}}(\hat{\mathbf{v}}_2(1), \hat{\mathbf{v}}_2(2), \ldots, \hat{\mathbf{v}}_2(n))^T$, where $\hat{\mathbf{v}}_2$ is the unit eigenvector corresponding to $\hat{\lambda}_2$, is called *normalised Fiedler vector* (c.f. Definition 1.2.15) and is used for clustering the graph G (see §1.2.3[1] and (Higham et al., 2007)). It is easy to see that

$$D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2 = \left(\frac{1}{\sqrt{D_G(1,1)}}\hat{\mathbf{v}}_2(1), \frac{1}{\sqrt{D_G(2,2)}}\hat{\mathbf{v}}_2(2), \ldots, \frac{1}{\sqrt{D_G(n,n)}}\hat{\mathbf{v}}_2(n)\right)^T.$$

Let the disjoint sets $\hat{V}_G^{(1)}$ and $\hat{V}_G^{(2)}$ be defined in the following way:

$$\hat{V}_G^{(1)} := \{v_i \mid \hat{\mathbf{v}}_2(i) \leq 0\} \qquad \text{and} \qquad \hat{V}_G^{(2)} := V_G \setminus \hat{V}_G^{(2)}. \tag{6.4}$$

Then it is known, as in the Property 5 of the Laplacian matrix, that the quantity

$$\sum_{v_i \in \hat{V}_G^{(1)}, v_j \in \hat{V}_G^{(2)}} w(v_i, v_j)$$

is small and the values of

$$\sum_{v_i \in \hat{V}_G^{(1)}, v_j \in \hat{V}_G^{(1)}} w(v_i, v_j) \qquad \text{and} \qquad \sum_{v_i \in \hat{V}_G^{(2)}, v_j \in \hat{V}_G^{(2)}} w(v_i, v_j)$$

are large.

Remark 6.1.1. *In Property 5 of the Normalised Laplacian matrix and in Property 5 of the Laplacian matrix the clusterings of G given by* (6.2) *and* (6.4) *are not the only possible ways to cluster G with respect to $\mathbf{v}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$. In fact, (Higham et al., 2007)) suggests clustering the graph by searching for significant gaps among the ordered elements of $\mathbf{v}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$, and then splitting the graph into two in such a way that the vertices, corresponding to the entries before the gap go in one cluster and the rest form the other cluster. In this respect it is probably worth noting that the order of the elements in $\hat{\mathbf{v}}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$ is the same. Therefore, in particular, when we cluster G with respect to the sign of the elements of $\hat{\mathbf{v}}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$, the clusters produced by these two vectors will be identical. However, in general, the gaps among the entries of $\hat{\mathbf{v}}_2$ and those among the entries of $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$ will be different (c.f. (Higham et al., 2007)).*

Let us summarise Property 5 of the Laplacian matrix and Property 5 of the normalised Laplacian matrix. The vectors $\mathbf{v}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$ respectively, are used to split the vertices of the graph G into two disjoint sets in such a way that the total weight of

[1]There we give a brief overview of some of the most popular approaches for graph clustering.

the edges between different clusters is small and the total weights of the edges, joining vertices from the same cluster, is large. So both vectors, $\mathbf{v}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$, seem to solve the same problem (which we haven't stated yet mathematically, but, as we shall see in the next paragraph, we don't need a rigorous statement). Thus, the question arises whether the solutions, produced by these two vectors, will differ in any way and if yes, how exactly will they differ. Another question worth asking is whether there exist classes of graphs, for which $\mathbf{v}_2$ and $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$ produce "similar" or "totally different" clusterings, although one first has to define mathematically the meaning of "similar" and "totally different" clusterings. Some authors (c.f. (Higham et al., 2007)) suggest that using $D_G^{-\frac{1}{2}}\hat{\mathbf{v}}_2$, instead of $\mathbf{v}_2$, makes the clustering "less susceptible to the influence of "poorly calibrated" vertices that have abnormally large or small weights" (quote from (Higham et al., 2007)).

We now reduce the questions raised above to comparing the signs of the entries of $\mathbf{v}_2$ and $\hat{\mathbf{v}}_2$. This is why we didn't need to state the problem of clustering mathematically. We shall simply imagine that the method of clustering of the graph G is the simplest, that which assigns vertices to clusters according to the sign of the corresponding entry in the *Fiedler vector* or in the *normalised Fiedler vector*. So, by comparing the signs of $\mathbf{v}_2$ and $\hat{\mathbf{v}}_2$, we shall be able to indicate the discrepancies between the clusterings produced by these two vectors.

In order to link $\mathbf{v}_2$ with $\hat{\mathbf{v}}_2$, and hence compare their entries, we create a Homotopy between the two eigenvalue problems,

$$L_G\mathbf{v} = \lambda\mathbf{v} \qquad \text{and} \qquad \hat{L}_G\hat{\mathbf{v}} = \hat{\lambda}\hat{\mathbf{v}}.$$

Let $t \in [0,1]$. Our aim is to find the second smallest eigenvalue, $\mu_2(t)$, of the generalised eigenvalue problem

$$L_G\mathbf{x}(t) = \mu(t)D_G(t)\mathbf{x}(t), \tag{6.5}$$

and its corresponding eigenvector of unit length, $\mathbf{x}_2(t)$, where

$$D_G(t) := tD_G + (1-t)I_n,$$

and compare the signs of the entries of $\mathbf{x}_2(0)$ and $\mathbf{x}_2(1)$.

In order to explain why problem (6.5) is helpful in addressing the questions above, let us consider the cases $t = 0$ and $t = 1$. For $t = 0$ we can see from (6.5) that we have to solve

$$L_G\mathbf{x}(0) = \mu(0)\mathbf{x}(0) \tag{6.6}$$

and find the second smallest eigenvalue, $\mu_2(0)$, and its corresponding eigenvector, $\mathbf{x}_2(0)$. Therefore

$$\lambda_2 = \mu_2(0) \qquad \text{and} \qquad \mathbf{v}_2 = \mathbf{x}_2(0).$$

When $t = 1$ we have to solve

$$L_G\mathbf{x}(1) = \mu(1)D_G\mathbf{x}(1), \tag{6.7}$$

which is equivalent to solving

$$D_G^{-\frac{1}{2}}L_GD_G^{-\frac{1}{2}}\mathbf{y} = \mu\mathbf{y}, \tag{6.8}$$

where $\mathbf{y} := D_G^{\frac{1}{2}}\mathbf{x}(1)$ and (6.8) is in fact identical to finding the eigenvalues and eigenvectors of $\hat{L}$. Hence,

$$\hat{\lambda}_2 = \mu_2(1) \qquad \text{and} \qquad \hat{\mathbf{v}}_2 = \mathbf{y}_2 = D_G^{\frac{1}{2}}\mathbf{x}_2(1).$$

The idea of introducing the Homotopy between $\mathbf{v}_2$ and $\hat{\mathbf{v}}_2$, or equivalently, between $\mathbf{x}_2(0)$ and $\mathbf{x}_2(1)$, is to find the law by which the signs of the entries of $\mathbf{x}_2(t)$ change with t. This would then show us the differences and similarities between the clusterings produced by $\mathbf{v}_2$ and $\hat{\mathbf{v}}_2$.

In this chapter we try to answer the question of whether clustering using L_G would be identical to clustering using $\hat{L}_G$ for products of Path graphs (see §6.3 for definitions and preliminary results and §6.4 for the main result). Even in this case, we were unable to answer that question. However, we have obtained the answer to a related question about clustering.

Finally, we mention a third matrix, which is usually associated with graphs, and its second smallest eigenvector is used for clustering in a similar way to $\mathbf{v}_2$ and $\hat{\mathbf{v}}_2$. This is the matrix $D_G^{-1}L_G$, where L_G is the Laplacian matrix of the graph. It is easy to see that the diagonal entries of that matrix are all equal to one and the off diagonal entries are equal to $-\frac{w(v_i,v_j)}{D_G(i,i)}$, $1 \leq i \neq j \leq n$, making the sum of the absolute values of its elements across each row is equal to 2. It is also easy to note that, in general, $D_G^{-1}L_G$ is not a symmetric matrix. However, the eigenvalue problem

$$D_G^{-1}L_G\mathbf{w} = \nu\mathbf{w} \qquad \text{is equivalent to} \qquad L_G\mathbf{w} = \nu D_G\mathbf{w}$$

and hence it is equivalent to

$$D_G^{-\frac{1}{2}}L_GD_G^{-\frac{1}{2}}\hat{\mathbf{v}} = \hat{\lambda}\hat{\mathbf{v}}$$

by letting $\nu := \hat{\lambda}$ and $\mathbf{w} := D_G^{-\frac{1}{2}}\hat{\mathbf{v}}$. Therefore all eigenvalues and eigenvectors of $D_G^{-1}L_G$ are real. Moreover, clustering with respect to the signs of the entries of $\mathbf{w}_2$ and $\hat{\mathbf{v}}_2$ would produce identical results. Hence, it is enough to compare only $\mathbf{v}_2$ with $\hat{\mathbf{v}}_2$.

6.2. Path graphs

In this section we recall the definition of Path graph (see §6.2.1) and some results about the spectral properties of symmetric and symmetric tridiagonal matrices (see §6.2.3). In particular, in Theorem 6.2.1 we recall a well-known result about the spectra of the Laplacian matrices of unweighted Path graphs.

6.2.1. Definitions

In order to give the definition of a *Path graph*, we need the definition of a *tree*.

Definition 6.2.1 (**Tree**)**.** Let G be a connected graph of order n. We say that G is a *tree*, if it has exactly $n-1$ edges of non-zero weight.

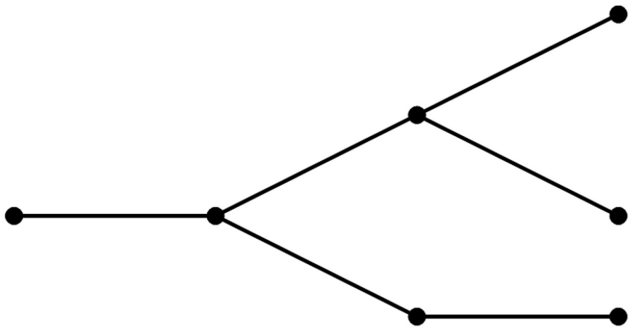

Figure 6-1: *An example of a* tree *with* 7 *vertices and* 6 *edges.*

Definition 6.2.2 (**Leaves**)**.** Let the graph $G = (V_G, E_G)$ be a tree. We say that the vertex $v \in V_G$ is a *leaf*, if it is adjacent to exactly one other vertex.

Definition 6.2.3 (**Path graph**)**.** *Path graph* is a tree of order $n \geq 2$ with exactly two leaves.

Figure 6-2: *An example of a* Path graph *with* 5 *vertices, joined by* 4 *edges.*

Remark 6.2.1. *It can be shown (e.g. by induction) that if $G = (V_G, E_G)$ is a Path graph, then the vertices in G, which are not leaves, are adjacent to exactly two other vertices in G. From this it can be further shown that there is a permutation of the elements in $V_G = \{v_1, v_2, \ldots, v_n\}$, so that the vertices v_1 and v_n are the leaves and the set of edges, E_G, is given by $E_G = \{\{v_i, v_{i+1}\} \mid 1 \leq i \leq n-1\}$.*

In this chapter Path graphs of order n shall usually be denoted by P_n.

6.2.2. Spectra of Unweighted Path graphs

The following result is well-known and is about the spectra of Laplacian matrices of Path graphs. It gives us an idea of how unweighted Path graphs are clustered with respect to their Laplacian matrices. In fact, it confirms what one would expect in that case. Namely, unweighted Path graphs are clustered by splitting them into two "equal" parts.

Theorem 6.2.1. *Let G be a path graph of order n with a weight function $w : E_G \to \{0, 1\}$ defined by $w(u, v) = 1$ for all $\{u, v\} \in E_G$ and $w(u, v) = 0$ for $\{u, v\} \notin E_G$. Further let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the matrix L_G and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be the eigenvectors corresponding to these eigenvalues. Then*

$$\lambda_k = 2 + 2\cos\left(\frac{(n-k+1)\pi}{n}\right),$$

where $k = 1, 2, \ldots, n$ and

$$\mathbf{v}_k(j) = \sin\left(\frac{k(2j-1)\pi}{2n}\right)$$

for $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n-1$ and

$$\mathbf{v}_n(j) = (-1)^{j-1}$$

for $j = 1, 2, \ldots, n$.

Proof. See (Yueh, 2005) for details. □

In §6.3.4 we prove an analogical result about the clustering with respect to the second eigenvector of the Laplacian matrix of a weighted Path graph (c.f. Lemma 6.3.3, in the case when $\xi = 0$).

6.2.3. Background results for symmetric and symmetric tridiagonal matrices

The following result is well-known and its proof omitted.

Theorem 6.2.2 (Weyl)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ be their eigenvalues, respectively. Let also $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ be the eigenvalues of $A + B$. Then*

$$\lambda_i + \beta_1 \leq \mu_i \leq \lambda_i + \beta_n \tag{6.9}$$

for $i = 1, 2, \ldots, n$.

Proof. See (Horn and Johnson, 1990), pp 181-182, for a proof. □

Corollary 6.2.1. *Let, in the settings of Theorem 6.2.2, B be a positive (negative) definite matrix. Then*

$$\lambda_i < \mu_i \qquad (\lambda_i > \mu_i) \tag{6.10}$$

for $i = 1, 2, \ldots, n$.

Before we state the next result, the Sylvester's law of inertia, we need the following definitions.

Definition 6.2.4. (**Inertia of a symmetric matrix**) Let A be a symmetric matrix and π, ν and ζ be the number of its positive, negative and zero eigenvalues, respectively. Then the triple (π, ν, ζ) is called A's *inertia*.

Definition 6.2.5. Let A and B be two symmetric matrices and (π_A, ν_A, ζ_A) and (π_B, ν_B, ζ_B) be their *inertias*, respectively. We say that the matrix A has the same *inertia* as B if $\pi_A = \pi_B$, $\nu_A = \nu_B$ and $\zeta_A = \zeta_B$.

Theorem 6.2.3 (**Sylvester's law of inertia**)**.** *Let A be a symmetric matrix and let B be a non-singular matrix. Then the matrix $B^T AB$ has the same inertia as A.*

Proof. This is a standard result in Matrix Theory (see (Horn and Johnson, 1990), p. 223, or (Spielman, 2004), Lecture 3.3). □

We now state results about tridiagonal matrices, which are helpful in generalising Theorem 6.2.1 for weighted Path graphs. These results will also be used in §6.4, where we state and prove our main result.

Theorem 6.2.4. *Every tridiagonal matrix with nonzero off-diagonal elements has simple spectrum.*

Proof. Let $A \in \mathbb{R}^{n\times n}$ be a tridiagonal matrix. Then for all $\lambda \in \mathbb{R}$ the minor of the element $(n, 1)$ in the matrix $A-\lambda I$ is $A_{1,2}A_{2,3}\cdots A_{n-1,n} \neq 0$. Therefore $rank(A-\lambda I) \geq n-1$ for all $\lambda \in \mathbb{R}$. □

Remark 6.2.2. *Theorem 6.2.4 implies that Laplacian matrices of Path graphs have a simple spectrum and, in particular, their second eigenvalue is simple. Thus, the Fiedler vector of a Path graph is determined uniquely, up to a scaling. Hence, clustering with respect to the entries of that vector is also unique.*

Lemma 6.2.1 (c.f. (Spielman, 2004), Lemma 3.3.3.)**.** *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric tridiagonal matrix with $2p$ positive off-diagonal entries such that*

$$M\mathbf{1} = \mathbf{0}.$$

Then M has p negative eigenvalues ($n > p$).

Remark 6.2.3. *The condition $M\mathbf{1} = \mathbf{0}$ describes a class of matrices in which the Laplacian matrices are only a subclass, since Laplacian matrices, as we consider them here, are required to have negative off-diagonal elements, because of their interpretation in Graph Theory.*

Proof. From Property 1 of Laplacian matrices (see Properties 6.1.1 above) we have

$$\mathbf{x}^T M\mathbf{x} = -\sum_{i=1}^{n-1} M_{i,i+1}(x_i - x_{i+1})^2$$

for any $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$. We now apply a change of variables from $\mathbf{x}$ to $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^T$ in the following way

$$x_i = \delta_1 + \delta_2 + \cdots + \delta_i$$

for $1 \leq i \leq n$. This change of variables is realised by the lower-triangular matrix L, which has 1's on and below the diagonal:

$$\mathbf{x} = L\boldsymbol{\delta}.$$

From Sylvester's inertia law (c.f. Theorem 6.2.3) we know that

$$L^T ML$$

has the same number of positive, negative and zero eigenvalues as M. On the other hand

$$\boldsymbol{\delta}^T L^T ML\boldsymbol{\delta} = -\sum_{i=1}^{n-1} M_{i,i+1}\delta_{i+1}^2,$$

so the matrix $L^T ML$ has one zero eigenvalue and as many negative eigenvalues, as

there are positive $M_{i,i+1}$. □

Theorem 6.2.5. *(c.f. (Spielman, 2004), Theorem 3.3.1.) Let $A \in \mathbb{R}^{n\times n}$ be a symmetric tridiagonal matrix with negative off-diagonal elements and let $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ be its eigenvalues. Further, let $\mathbf{v}_k$ be an eigenvector corresponding to λ_k, $1 \leq k \leq n$. Then $\mathbf{v}_k$ changes sign $k-1$ times.*

Proof. We will just consider the case in which $\mathbf{v}_k$ has no zero entries. In this case we wish to show that the number of i's, for which $\mathbf{v}_k(i)\mathbf{v}_k(i+1) < 0$, equals $k-1$, where $\mathbf{v}_k(i)$ is the i-th entry of the vector $\mathbf{v}_k$.

Let us define the diagonal matrix V_k in the following way:

$$V_k(i,i) = \mathbf{v}_k(i)$$

for $1 \leq i \leq n$. Consider the matrix

$$M := V_k^T(A - \lambda_k I)V_k. \tag{6.11}$$

Since $A-\lambda_k I$ has $k-1$ negative eigenvalues, by Sylvester's law of inertia, Theorem 6.2.3, M has $k-1$ negative eigenvalues, one zero eigenvalue and $n-k$ positive eigenvalues. Note that

$$M\mathbf{1} = \mathbf{0} \qquad \text{and} \qquad M_{i,i+1} = \mathbf{v}_k(i)A_{i,i+1}\mathbf{v}_k(i+1).$$

Therefore $M_{i,i+1} > 0$, if and only if $\mathbf{v}_k(i)\mathbf{v}_k(i+1) < 0$. Thus, by Lemma 6.2.1, there are exactly $k-1$ such i's. □

6.3. Cartesian products of Path Graphs

In this section we recall the definition of Cartesian products of graphs and the definitions of consistent and non-consistent weight functions. We further recall the definition of a weight function on Cartesian products of graphs and construct a non-consistent weight function on the product of Path graphs. The latter is shown to lead to the derivation of the Laplacian matrix of the product of Path graphs as a Kronecker product, which involves the Laplacian matrices of the respective "one-dimensional" Path graphs. Thus, the basic properties of the Kronecker product are recalled. Then, the Homotopy between the Laplacian and the normalised Laplacian matrices of products of Path graphs is given. Finally, that Homotopy problem is shown to be equivalent to a generalised eigenvalue problem, which can be split into two smaller problems, each involving only the Laplacian matrices of the "one-dimensional" Path graphs.

6.3.1. Cartesian products of graphs

Here we define the Cartesian product of two (or more) graphs in general. The definitions of this subsection are used later, when we restrict our attention to products of Path graphs only.

Definition 6.3.1. Let $G = (V_G, E_G)$ and $H = (V_H, E_H)$ be two graphs. Then $G \times H$ is the graph with vertex set

$$V_{G\times H} = V_G \times V_H = \{(v, w) \mid v \in V_G \text{ and } w \in V_H\}$$

and edge set

$$E_{G\times H} = \{\{(v, w_1), (v, w_2)\} \mid \{w_1, w_2\} \in E_H\} \cup \{\{(v_1, w), (v_2, w)\} \mid \{v_1, v_2\} \in E_G\}$$

(see Figure 6-3 and 6-4).

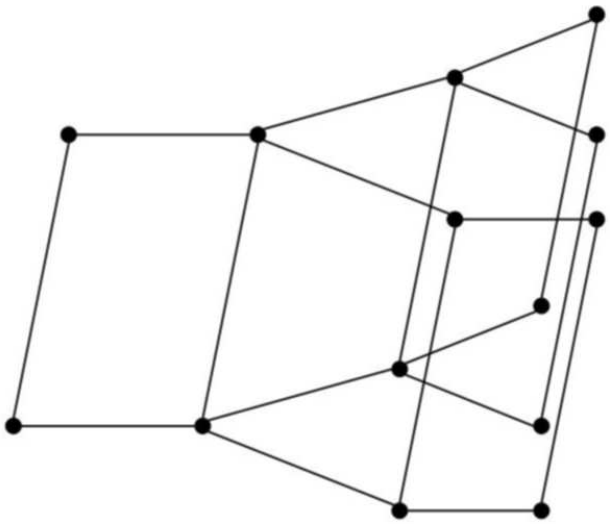

Figure 6-3: *Cartesian product of a* tree *and a* Path graph.

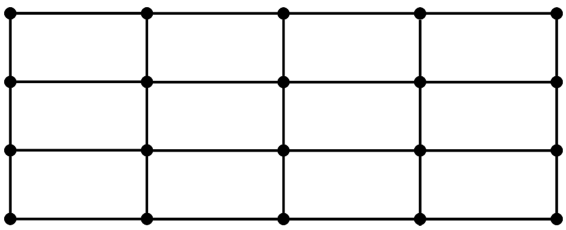

Figure 6-4: *Cartesian product of two* Path graphs, $P_5 \times P_4$.

Let us recall Definitions 1.2.7 and 1.2.8, and Remark 1.2.3 from §1.

Definition 6.3.2. Let $G = (V_G, E_G)$ be a graph and w be some weight function on G. We say that w is *consistent*, if

$$w(v) = \sum_{u \in V_G} w(u, v). \tag{6.12}$$

Remark 6.3.1. *When w is a consistent weight function, the quantity $w(v)$, $v \in V_G$, is usually called the* degree *of vertex v and is denoted by d_v.*

Definition 6.3.3. A weight function which is not consistent is called a *non-consistent* weight function.

Definition 6.3.4. Let G and H be two graphs with weight functions w_G and w_H, respectively (w_G and w_H not necessarily being *consistent* weight functions). The weight function $w_{G\times H}$ on the product graph $G \times H$ is defined in the following way:

$$w_{G\times H}\big((u,v),(u,v')\big) = w_G(u)w_H(v,v') \quad \text{and} \quad w_{G\times H}\big((u,v),(u',v)\big) = w_G(u,u')w_H(v).$$

6.3.2. Non-consistent weight function and the Kronecker product

In this chapter we consider only non-consistent weight functions on Path graphs, which are used in the construction of corresponding weight functions on the Cartesian product of Path graphs. We only briefly mention here that consistent weight functions correspond naturally to random walks on graphs (c.f. (Chung, 1997), p. 37).

Proposition 6.3.1. *Let $P_n = \{v_1, v_2, \ldots, v_n\}$, $n \in \mathbb{N}$, be a Path Graph with a weight function w defined by*

$$w(v_i, v_{i+1}) > 0 \quad \textit{for } 1 \le i \le n-1 \qquad \textit{and} \qquad w(v_i) = 1 \quad \textit{for } 1 \le i \le n. \tag{6.13}$$

Then w is non-consistent *for $n > 1$.*

Proof. Let us assume that w is *consistent*. For the quantities $w(v_1)$ and $w(v_2)$ we have

$$1 = w(v_1) = w(v_1, v_2)$$

and therefore

$$1 = w(v_2) = w(v_1, v_2) + w(v_2, v_3) = 1 + w(v_2, v_3)$$

which implies $w(v_2, v_3) = 0$. A contradiction. Therefore w_G is *non-consistent*. □

Let $G = P_m = \{u_1, u_2, \ldots, u_m\}$ and $H = P_n = \{v_1, v_2, \ldots, v_n\}$, $m, n \in \mathbb{N}$, be two Path Graphs and each of their weight functions, w_G and w_H respectively, be defined as in (6.13). Then, according to Definition 6.3.4, we can define a weight function, $w_{G\times H}$, on the cartesian product of G and H in the following way:

$$w_{G\times H}\big((u_i, v_j), (u_i, v_k)\big) := w_H(v_j, v_k) \qquad \text{and} \qquad w_{G\times H}\big((u_j, v_i), (u_k, v_i)\big) := w_G(u_j, u_k) \tag{6.14}$$

for all $1 \leq i \leq n$ and all $1 \leq j < k \leq n$. Also, let us define $w_{G\times H}$ on the vertices of $G \times H$, (u_i, v_j), $1 \leq i, j \leq n$, as follows:

$$w_{G\times H}\big((u_i, v_j)\big) := \sum_{u_k \sim u_i} w_G(u_k, u_i) + \sum_{v_l \sim v_j} w_H(v_l, v_j). \tag{6.15}$$

Then the weight function $w_{G\times H}$, defined by (6.14) and (6.15), is a consistent weight function. Before we state Proposition 6.3.2, we recall some standard notation.

Definition 6.3.5. (**Kronecker product**) Let $A = (a_{ij})_{1\leq i,j\leq m} \in \mathbb{R}^{m\times m}$ and $B = (b_{ij})_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$ be two matrices. Then the $nm \times nm$ matrix

$$A \otimes B := (a_{ij}B)_{1\leq i,j\leq m}$$

is called the (right) Kronecker product of A and B.

We present some well-known properties of the Kronecker product (c.f. (McDuffee, 1946)).

Properties 6.3.1. (Kronecker product)

1. For any three matrices, $A_1, A_2 \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$, the following holds

$$(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B.$$

2. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$.

3. $(A \otimes B)^T = A^T \otimes B^T$.

4. Let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$ be two matrices and $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ be two column vectors. Then

$$(A \otimes B)(\mathbf{x} \otimes \mathbf{y}) = (A\mathbf{x}) \otimes (B\mathbf{y}).$$

5. Let $\varphi(\xi,\eta)=\sum_{i,j}c_{ij}\xi^i\eta^j$ be a polynomial in ξ and η, and define

$$\varphi(A;B)=\sum_{i,j}c_{ij}A^i\otimes B^j,$$

then the eigenvalues of $\varphi(A;B)$ are the functions $\varphi(\lambda_A,\lambda_B)$, where λ_A and λ_B independently take all possible eigenvalues of the matrices A and B, respectively.

The following proposition is a standard result in (Spectral) Graph Theory.

Proposition 6.3.2. *Let $G=P_m=\{u_1,u_2,\ldots,u_m\}$ and $H=P_n=\{v_1,v_2,\ldots,v_n\}$, $m,n\in\mathbb{N}$, be two Path Graphs and the weight function $w_{G\times H}$ be defined by* (6.14) *and* (6.15)*. Then the Laplacian matrix of the graph $G\times H$ satisfies*

$$L_{G\times H}=L_G\otimes I_n+I_m\otimes L_H, \tag{6.16}$$

where the symbol $\otimes$ is used to denote the (right) Kronecker product of two matrices and L_G and L_H are the Laplacian matrices of G and H, respectively.

It is easy to show, from Property 3 of the Kronecker product, that $L_{G\times H}$ is a symmetric matrix, since

$$L_{G\times H}^T=(L_G\otimes I_m)^T+(I_n\otimes L_H)^T=L_G^T\otimes I_n^T+I_m^T\otimes L_H^T=L_{G\times H}, \tag{6.17}$$

where we have also used that L_G and L_H are symmetric matrices.

Also, if $\mathbf{1}_m$, $\mathbf{1}_n$ and $\mathbf{1}_{nm}$ are the column vectors whose entries are all equal to one in $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^{nm}$ respectively, then $\mathbf{1}_m\otimes\mathbf{1}_n=\mathbf{1}_{nm}$ and

$$\begin{aligned} L_{G\times H}\mathbf{1}_{nm} &= (L_G\otimes I_n+I_m\otimes L_H)(\mathbf{1}_m\otimes\mathbf{1}_n) \\ &= (L_G\mathbf{1}_m)\otimes\mathbf{1}_n+\mathbf{1}_m\otimes(L_H\mathbf{1}_n)=\mathbf{0}_m\otimes\mathbf{0}_n=\mathbf{0}_{nm}, \end{aligned} \tag{6.18}$$

where $\mathbf{0}_m$, $\mathbf{0}_n$ and $\mathbf{0}_{nm}$ are the vectors whose entries are all zeros in $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^{nm}$, respectively. In the derivation of (6.18) we have used Properties 1 and 4 of the Kronecker product and the fact that L_G and L_H are Laplacian matrices.

Therefore, from (6.17) and (6.18) it follows that $L_{G\times H}$ is indeed a symmetric Laplacian matrix.

Let us denote by $D_{G\times H}$ the matrix containing the diagonal elements of $L_{G\times H}$, whose off-diagonal entries are all zero. We recall that the corresponding matrices for L_G and L_H were D_G and D_H, respectively. Then one can see from (6.16) that

$$D_{G\times H}=D_G\otimes I_m+I_n\otimes D_H. \tag{6.19}$$

The discussion above is summarised in the following corollary.

Corollary 6.3.1. *Let $G = P_m = \{u_1, u_2, \ldots, u_m\}$ and $H = P_n = \{v_1, v_2, \ldots, v_n\}$, $m, n \in \mathbb{N}$, be two Path Graphs and the weight function $w_{G\times H}$ be defined by* (6.14) *and* (6.15)*. Then the normalised Laplacian matrix of the graph $G \times H$, which we denote by $\hat{L}_{G\times H}$, is given by*

$$\hat{L}_{G\times H} = D_{G\times H}^{-\frac{1}{2}} L_{G\times H} D_{G\times H}^{-\frac{1}{2}},$$

where the matrices $L_{G\times H}$ and $D_{G\times H}$ are given by (6.16) *and* (6.19)*, respectively.*

6.3.3. Homotopy between the normalised and the unnormalised Laplacian matrices of Cartesian products of graphs

In this subsection we write the Homotopy problem, introduced in (6.5), in terms of the Laplacian matrix of the Cartesian product of the graphs G and H. Then we make an equivalent statement of that problem, which splits it into two similar problems of smaller dimension, each involving only L_G and D_G, and L_H and D_H, respectively. The main advantage of splitting the bigger Homotopy problem into two problems of smaller size, is that each of the latter involves only tridiagonal matrices, when G and H are Path graphs. Thus, we can apply the results from §6.2.3 to each of the "smaller" problems and in this way we can derive results about the initial Homotopy problem.

The Homotopy problem. Problem (6.5) in terms of the Laplacian matrix of the product $G \times H$, $L_{G\times H}$, becomes

$$\big[L_G \otimes I_n + I_m \otimes L_H\big]\mathbf{x}(t) = \mu(t)[D_G(t) \otimes I_n + I_m \otimes D_H(t)]\mathbf{x}(t), \tag{6.20}$$

where $t \in [0, 1]$, $D_G(t) = tD_G + (1-t)I_m$ and $D_H(t) = tD_H + (1-t)I_n$. Since

$$L_{G\times H} = L_G \otimes I_n + I_m \otimes L_H$$

is a Laplacian and thus positive semi-definite matrix, and the diagonal matrix

$$D_{G\times H} = D_G \otimes I_n + I_m \otimes D_H$$

contains only positive elements on its main diagonal, one can easily show that for each $t \in [0, 1]$ the generalised eigenvalue problem (6.20) has nm eigenvalues,

$$0 = \mu_1(t) < \mu_2(t) \leq \mu_3(t) \leq \cdots \leq \mu_{nm}(t). \tag{6.21}$$

Here we have used the fact that the graph $G \times H$ is connected and therefore $\mu_2(t) > 0$.

An equivalent problem. Equation (6.20) can be rewritten in the following form:

$$\big[(L_G - \mu(t)D_G(t)) \otimes I_n + I_m \otimes (L_H - \mu(t)D_H(t))\big]\mathbf{x}(t) = 0. \tag{6.22}$$

Let $\xi \in \mathbb{R}$ and consider a different problem, that of finding $\theta(\xi, t) \in \mathbb{R}$ and $\mathbf{z}(\xi, t) \in \mathbb{R}^{nm}$ satisfying

$$\big[(L_G - \xi D_G(t)) \otimes I_n + I_m \otimes (L_H - \xi D_H(t))\big]\mathbf{z}(\xi, t) = \theta(\xi, t)\mathbf{z}(\xi, t). \tag{6.23}$$

For each $\xi \in \mathbb{R}$ and $t \in [0, 1]$ (6.23) is a symmetric eigenvalue problem and therefore there are nm real eigenvalues,

$$\theta_1(\xi, t) \leq \theta_2(\xi, t) \leq \cdots \leq \theta_{nm}(\xi, t)$$

and nm real eigenvectors corresponding to them,

$$\mathbf{z}_1(\xi, t), \mathbf{z}_2(\xi, t), \ldots, \mathbf{z}_{nm}(\xi, t).$$

The relation between (6.23) and (6.22), and thus between (6.23) and (6.20), is that when t is fixed and ξ_0 is such that

$$\theta_{i_0}(\xi_0, t) = 0 \quad \text{for some } 1 \leq i_0 \leq nm,$$

we obtain

$$\big[(L_G - \xi_0 D_G(t)) \otimes I_n + I_m \otimes (L_H - \xi_0 D_H(t))\big]\mathbf{z}_{i_0}(\xi_0, t) = 0$$

and therefore

$$\big[L_G \otimes I_n + I_m \otimes L_H\big]\mathbf{z}_{i_0}(\xi_0, t) = \xi_0\big[D_G(t) \otimes I_n + I_m \otimes D_H(t)\big]\mathbf{z}_{i_0}(\xi_0, t).$$

Hence

$$\mu_i(t) = \xi_0 \qquad \text{and} \qquad \mathbf{x}_i(t) = \mathbf{z}_{i_0}(\xi_0, t)$$

for some $1 \leq i \leq nm$.

We now show that the converse is also true. Let the eigenpair $\mu_i(t)$ and $\mathbf{x}_i(t)$ for some $1 \leq i \leq nm$ be a solution of (6.20), and thus also of (6.22). Then from (6.22) we have

$$\big[(L_G - \mu_i(t)D_G(t)) \otimes I_n + I_m \otimes (L_H - \mu_i(t)D_H(t))\big]\mathbf{x}_i(t) = 0 \cdot \mathbf{x}_i(t).$$

Therefore, if we let $\xi_0 := \mu_i(t)$ in (6.23), then there exists a number $1 \leq i_0 \leq nm$ such that

$$\theta_{i_0}(\xi_0, t) = 0 \qquad \text{and} \qquad \mathbf{z}_{i_0}(\xi_0, t) = \mathbf{x}_i(t).$$

Moreover, we can also show that if for some $1 \leq i \leq nm$ $\mu_i(t)$ is an eigenvalue of multiplicity k ($k \geq 1$) in (6.20), then the multiplicity of the eigenvalue 0 in

$$\big[(L_G - \mu_i(t)D_G(t)) \otimes I_n + I_m \otimes (L_H - \mu_i(t)D_H(t))\big]\mathbf{z}(\xi, t) = \theta(\xi, t)\mathbf{z}(\xi, t)$$

is also equal to k.

Hence, for $t \in [0, 1]$ fixed, the number of all different ξ's, for which at least one of the eigenvalues, $\theta_i(\xi, t)$, in (6.23) is equal to zero, is equal to the number of distinct eigenvalues in (6.20) (or, equivalently, in (6.22)). Further, if each such ξ is counted as many times as is the multiplicity of the eigenvalue 0 in (6.23) at that particular ξ, then the total number of all such ξ's is nm. Therefore, if we put all such ξ's in an increasing order, we have

$$\xi_1 \leq \xi_2 \leq \cdots \leq \xi_{nm} \qquad \text{and} \qquad \mu_i(t) = \xi_i, \quad 1 \leq i \leq nm. \tag{6.24}$$

Hence, using (6.21), we obtain

$$0 = \xi_1 < \xi_2 \leq \xi_3 \leq \cdots \leq \xi_{nm}. \tag{6.25}$$

The reason for considering (6.23) instead of (6.20), or (6.22), is that in (6.23) we can use Property 5 of the Kronecker product. We have explained this more precisely in the next theorem.

Theorem 6.3.1. *Let $t \in [0, 1]$, $\xi \in \mathbb{R}$ and let the eigenvalues of $L_G - \xi D_G(t)$ be*

$$\theta_{G1}(\xi, t) \leq \theta_{G2}(\xi, t) \leq \cdots \leq \theta_{Gm}(\xi, t)$$

and $\mathbf{z}_{Gi}(\xi, t)$, $1 \leq i \leq m$, be their corresponding unit eigenvectors. Similarly, let

$$\theta_{H1}(\xi, t) \leq \theta_{H2}(\xi, t) \leq \cdots \leq \theta_{Hn}(\xi, t)$$

be the eigenvalues of $L_H - \xi D_G(t)$ and $\mathbf{z}_{Hj}(\xi, t)$, $1 \leq j \leq n$, be their corresponding unit eigenvectors. Then the eigenvalues and eigenvectors in (6.23)*, $\theta_s(\xi, t)$ and $\mathbf{z}_s(\xi, t)$,*

$1 \leq s \leq nm$, *satisfy*

$$\{\theta_s(\xi,t) \mid s = 1,2,\ldots,nm\} = \{\theta_{Gi}(\xi,t)+\theta_{Hj}(\xi,t) \mid i = 1,2,\ldots,m \text{ and } j = 1,2,\ldots,n\}, \tag{6.26}$$

and

$$\{\mathbf{z}_s(\xi,t) \mid s = 1,2,\ldots,nm\} = \{\mathbf{z}_{Gi}(\xi,t)\otimes\mathbf{z}_{Hj}(\xi,t) \mid i = 1,2,\ldots,m \text{ and } j = 1,2,\ldots,n\}. \tag{6.27}$$

Proof. The proof of this theorem is a simple application of Property 5 of the Kronecker product (see Properties 6.3.1). □

Remark 6.3.2. *The meaning of* (6.26) *and* (6.27) *in the statement Theorem 6.3.1 is that each eigenvalue* $\theta_s(\xi,t)$ *is equal to* $\theta_{Gi}(\xi,t) + \theta_{Hj}(\xi,t)$*, for some* $1 \leq i \leq m$ *and some* $1 \leq j \leq n$*, and the eigenvector corresponding to* $\theta_s(\xi,t)$*,* $\mathbf{z}_s(\xi,t)$*, is equal to* $\mathbf{z}_{Gi}(\xi,t) \otimes \mathbf{z}_{Gj}(\xi,t)$*, for exactly the same pair of indices,* i *and* j*.*

The conclusion from Theorem 6.3.1, (6.25) and the preceding discussion is that, given a $t \in [0,1]$, in order to find $\mu_2(t)$ and its corresponding unit eigenvector in (6.20), $\mathbf{x}_2(t)$, we have to find the smallest $\xi > 0$, such that

$$0 \in \{\theta_{Gi}(\xi,t) + \theta_{Hj}(\xi,t) \mid i = 1,2,\ldots,m \text{ and } j = 1,2,\ldots,n\}. \tag{6.28}$$

Once we have found such a ξ if, for example, $\theta_{Gi_0}(\xi,t)$ and $\theta_{Hj_0}(\xi,t)$ are such that $\theta_{Gi_0}(\xi,t) + \theta_{Hj_0}(\xi,t) = 0$, then the vector $\mathbf{x}_2(\xi,t)$ will be given by

$$\mathbf{x}_2(\xi,t) = \mathbf{z}_{Gi_0}(\xi,t) \otimes \mathbf{z}_{Hj_0}(\xi,t),$$

where $\mathbf{z}_{Gi_0}(\xi,t)$ and $\mathbf{z}_{Hj_0}(\xi,t)$ are the eigenvectors of $L_G - \xi D_G(t)$ and $L_H - \xi D_H(t)$ corresponding to the eigenvalues $\theta_{Gi_0}(\xi,t)$ and $\theta_{Hj_0}(\xi,t)$, respectively. Assuming that in (6.25) $\xi_2 < \xi_3$, the smallest positive ξ, for which (6.28) is satisfied, will be unique and will be equal to ξ_2. Further, according to (6.24), we will have $\mu_2(t) = \xi_2$.

Therefore we have reduced the Homotopy problem (6.20) into finding the smallest positive ξ, for which (6.28) holds, where $\theta_{Gi}(\xi,t)$ and $\theta_{Hj}(\xi,t)$ are the eigenvalues of the matrices $L_G - \xi D_G(t)$ and $L_H - \xi D_H(t)$, respectively. In the next subsection we take a closer look at the spectra of these two matrices in the case when G and H are Path graphs. In doing so we use the results in §6.2.3, since $L_G - \xi D_G(t)$ and $L_H - \xi D_H(t)$ in that case are symmetric tridiagonal matrices.

6.3.4. The spectrum of the matrix $L_G - \xi D_G(t)$ when G is a Path graph

Following the discussion in the last subsection, here we investigate the behaviour of the eigenvalues and eigenvectors of the matrix $L_G - \xi D_G(t)$ in the case when G is a Path graph.

We start with the observation that when G is a Path graph of order $m \geq 2$, that is, $G = P_m$ for some $m \geq 2$, from Remark 6.2.1 it follows that its Laplacian, L_G, is a symmetric tridiagonal matrix with $L_G(i, i+1) = -w(v_i, v_{i+1})$, where we recall that $w(v_i, v_{i+1})$ is the weight of the edge between vertices v_i and v_{i+1}, $1 \leq i \leq m-1$. Therefore $L_G(i,i) = D_G(i,i) = w(v_i, v_{i-1}) + w(v_i, v_{i+1})$, $2 \leq i \leq m-1$, $L_G(1,1) = D_G(1,1) = w(v_1, v_2)$ and $L_G(m,m) = D_G(m,m) = w(v_{m-1}, v_m)$. Thus, the matrix $L_G - \xi D_G(t)$ will also be symmetric and tridiagonal. Hence, an immediate consequence of Theorem 6.2.4 is the following lemma.

Lemma 6.3.1. *Let the Path graph, $G = P_m$, be connected. Then the spectrum of the matrix*

$$L_G - \xi D_G(t)$$

is simple for all $\xi \in \mathbb{R}$ and all $t \in [0,1]$.

Proof. When the Path graph, P_m, is connected, the off-diagonal elements of its Laplacian matrix, L_G, are all negative and thus so are the off-diagonal entries of the matrix $L_G - \xi D_G(t)$ for all $\xi \in \mathbb{R}$ and all $t \in [0,1]$. Therefore we can apply Theorem 6.2.4 to the matrix $L_G - \xi D_G(t)$. □

Lemma 6.3.2. *Let the Path graph, $G = P_m$, be connected. Then, for every $t \in [0,1]$, each of the eigenvalues of the matrix $L_G - \xi D_G(t)$, $\theta_{Gi}(\xi, t_0)$, $1 \leq i \leq n$, is a decreasing function of ξ for $\xi \geq 0$.*

Proof. Let us fix $t \in [0,1]$. If $\xi > 0$, the matrix $\xi D_G(t)$ is positive definite when P_m is connected. Therefore, if $0 < \xi^{(1)} < \xi^{(2)}$, we have

$$L_G - \xi^{(1)} D_G(t) = L_G - \xi^{(2)} D_G(t) + \left(\xi^{(2)} - \xi^{(1)}\right) D_G(t),$$

where the matrix $\left(\xi^{(2)} - \xi^{(1)}\right) D_G(t)$ is positive definite. Hence, by Corollary 6.2.1, we obtain

$$\theta_{Gi}(\xi_2, t) < \theta_{Gi}(\xi_1, t)$$

for all $i = 1, 2, \ldots, m$. □

Lemma 6.3.3. *Let $G = P_m$ be a Path graph and L_G be its Laplacian matrix. Further, let*

$$\theta_{G1}(\xi,t) < \theta_{G2}(\xi,t) < \cdots < \theta_{Gm}(\xi,t)$$

be the eigenvalues of the matrix $L_G - \xi D_G(t)$ and let $\mathbf{z}_{G1}(\xi,t), \mathbf{z}_{G2}(\xi,t), \ldots, \mathbf{z}_{Gm}(\xi,t)$ be their corresponding unit eigenvectors. Then, for all $t \in [0,1]$ and all $\xi \in \mathbb{R}$, the eigenvector $\mathbf{z}_{Gi}(\xi,t)$ changes sign $i-1$ times.

Proof. When G is a Path graph, for every $t \in [0,1]$ and every $\xi \in \mathbb{R}$ the matrix $L_G - \xi D_G(t)$ is symmetric and tridiagonal. Therefore we can apply Theorem 6.2.5 to the eigenvector $\mathbf{z}_{Gi}(\xi,t)$ and conclude that it changes sign $i-1$ times. □

The last lemma doesn't specify where the changes of sign in the entries of the eigenvectors occur. It only tells us that they exist. It turns out that, when we consider unweighted Path graphs, we can tell exactly where the change of sign in the entries of $\mathbf{z}_{G1}(\xi,t)$ and $\mathbf{z}_{G2}(\xi,t)$ occurs. This is the result of the next corollary.

Corollary 6.3.2. *Let $G = P_m$ be an unweighted Path graph and let L_G be its Laplacian matrix. Further, let $\mathbf{z}_{G1}(\xi,t)$ and $\mathbf{z}_{G2}(\xi,t)$ be the smallest and the second smallest eigenvectors of $L_G - \xi D_G(t)$, respectively. Then, for all $t \in [0,1]$ and all $\xi \in \mathbb{R}$ we have*

$$\mathbf{z}_{G1}^{[j]}(\xi,t) = \mathbf{z}_{G1}^{[m-j]}(\xi,t) \qquad \text{and} \qquad \mathbf{z}_{G2}^{[j]} = -\mathbf{z}_{G2}^{[m-j]}(\xi,t),$$

for $1 \leq j \leq m$, where $\mathbf{z}_{Gi}^{[j]}(\xi,t)$ denotes the j-th entry of the vector $\mathbf{z}_{Gi}(\xi,t)$.

Proof. When G is an unweighted Path graph of order m, its Laplacian matrix, L_G, is given by

$$L_G = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}$$

and therefore the diagonal matrix, $D_G(t)$, satisfies

$$D_G(t) = t \begin{bmatrix} 1 & & & & & \\ & 2 & & & & \\ & & 2 & & & \\ & & & \ddots & & \\ & & & & 2 & \\ & & & & & 1 \end{bmatrix} + (1-t) \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix},$$

which implies

$$D_G(t) = (1+t) \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix} - t \begin{bmatrix} 1 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & \\ & & & & & 1 \end{bmatrix}.$$

Hence, if $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ is a vector of unit norm, the Rayleigh quotient

$$\begin{aligned} & \mathbf{x}^T(L_G - \xi D_G(t))\mathbf{x} = \mathbf{x}^T L_G \mathbf{x} + \xi t(x_1^2 + x_m^2) - (1+t)\xi \qquad (6.29) \\ = \;& (x_1 - x_2)^2 + (x_2 - x_3)^2 + \cdots + (x_{m-1} - x_m)^2 + \xi t(x_1^2 + x_m^2) - (1+t)\xi. \end{aligned}$$

Now let us suppose that $\mathbf{x} = \mathbf{z}_{G2}(\xi, t)$. Then, from (6.29),

$$(x_1-x_2)^2 + (x_2-x_3)^2 + \cdots + (x_{m-1}-x_m)^2 + \xi t(x_1^2+x_m^2) - (1+t)\xi = \theta_{G2}(\xi, t). \quad (6.30)$$

Let us define the vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ so that

$$y_j := x_{m-j}, \quad 1 \le j \le m.$$

Then it is easy to see from (6.30) that $\|\mathbf{y}\|_2 = 1$ and

$$(y_1 - y_2)^2 + (y_2 - y_3)^2 + \cdots + (y_{m-1} - y_m)^2 + \xi t(y_1^2 + y_m^2) - (1+t)\xi = \theta_{G2}(\xi, t).$$

Thus, since the eigenvalue $\theta_{G2}(\xi, t)$ is simple (see Lemma 6.3.1), we have $\mathbf{x} = \alpha\mathbf{y}$ for some $\alpha \in \mathbb{R}$. But since $\|\mathbf{y}\|_2 = 1$, we have $|\alpha| = 1$. Further, since by Lemma 6.3.3 the vector $\mathbf{z}_{G2}(\xi, t)$, and thus $\mathbf{x}$, changes sign once, we have $sign(x_1) = -sign(x_m)$ and

hence $\alpha = -1$. Therefore

$$x_j = -x_{m-j}, \quad 1 \leq j \leq m.$$

The proof that

$$\mathbf{z}_{G1}^{[j]}(\xi, t) = \mathbf{z}_{G1}^{[m-j]}(\xi, t), \quad 1 \leq j \leq m \tag{6.31}$$

is similar. The only difference is that, by Lemma 6.3.3, the eigenvector $\mathbf{z}_{G1}(\xi, t)$ doesn't change sign and therefore $\alpha = 1$ in this case. The latter implies (6.31). □

Remark 6.3.3. *The last corollary tells us that the entries of the eigenvector $\mathbf{z}_{G1}(\xi, t)$ are symmetric with respect to the number of its middle entry (or middle entries, if m is even) for all $t \in [0, 1]$ and all $\xi \in \mathbb{R}$. It also shows that, for all $t \in [0, 1]$ and all $\xi \in \mathbb{R}$, the change of sign in the eigenvector $\mathbf{z}_{G2}(\xi, t)$ occurs at the entry with number $\frac{m}{2}$. More precisely, when m is even, the two middle entries (i.e. those with numbers $\frac{m}{2}$ and $\frac{m}{2} + 1$) are equal in magnitude, but with opposite signs, and when m is odd, the middle entry (i.e. the one with number $\lfloor \frac{m}{2} \rfloor$) is equal to zero and this is where the change of sign occurs. (Here $\lfloor x \rfloor$ denotes the integer part of the real number x.)*

6.4. Main result

In this section we state and prove our main result. An immediate implication of it is that the two Laplacian matrices, $L_{G\times H}$ and $\hat{L}_{G\times H}$, introduced in §6.3.2, cluster any product of Path graphs, weighted or unweighted, by "cutting" them "horizontally" or "vertically".

We start with a simple corollary from Lemma 6.3.2.

Corollary 6.4.1. *Let $t \in [0, 1]$ be fixed. Then, using the notation of §6.3.3, the eigenvalues $\theta_s(\xi, t)$, $1 \leq s \leq nm$, in* (6.23) *are decreasing functions of ξ for all $\xi \geq 0$.*

Proof. Let $t \in [0, 1]$ be fixed. By Theorem 6.3.1 each eigenvalue $\theta_s(\xi, t)$ in (6.23) is given by

$$\theta_s(\xi, t) = \theta_{Gi}(\xi, t) + \theta_{Hj}(\xi, t)$$

for some pair of indices, i and j, where $1 \leq i \leq m$ and $1 \leq j \leq n$. By Lemma 6.3.2 all eigenvalues $\theta_{Gi}(\xi, t)$ and $\theta_{Hj}(\xi, t)$, where $1 \leq i \leq m$ and $1 \leq j \leq n$ are decreasing functions of ξ for all $\xi \geq 0$. Therefore $\theta_s(\xi, t)$ is also a decreasing function of ξ for all $\xi \geq 0$. □

Further, for any given $t \in [0, 1]$ and $\xi \in \mathbb{R}$, we know that $\theta_1(\xi, t) = \theta_{G1}(\xi, t) + \theta_{H1}(\xi, t)$, since

$$\theta_{G1}(\xi, t) + \theta_{H1}(\xi, t) < \theta_{Gi}(\xi, t) + \theta_{Hj}(\xi, t)$$

for all $i = 2, 3, \ldots, m$ and all $j = 2, 3, \ldots, n$. Therefore

$$0 = \theta_1(0,t) = \theta_{G1}(0,t) + \theta_H(0,t) < \theta_{Gi}(0,t) + \theta_{Hj}(0,t)$$

for all $i = 2, 3, \ldots, m$ and all $j = 2, 3, \ldots, n$. Also, in a similar way, $\theta_2(\xi, t)$ will be equal to

$$\theta_{G1}(\xi,t) + \theta_{H2}(\xi,t) \qquad \text{or} \qquad \theta_{G2}(\xi,t) + \theta_{H1}(\xi,t),$$

whichever is smaller. Hence, according to the definition of ξ_2 (see (6.25) and the preceding discussion), we have

$$\theta_{G1}(\xi_2,t) + \theta_{H2}(\xi_2,t) = 0 \qquad \text{or} \qquad \theta_{G2}(\xi_2,t) + \theta_{H1}(\xi_2,t) = 0.$$

From Corollary 6.4.1 and Lemma 6.3.2 at the point ξ_2 we have

$$\theta_1(\xi_2,t) < 0 = \theta_2(\xi_2,t) \leq \theta_3(\xi_2,t) \leq \cdots \leq \theta_{nm}(\xi_2,t).$$

Let us suppose that

$$0 = \theta_2(\xi_2,t) = \theta_{G1}(\xi_2,t) + \theta_{H2}(\xi_2,t).$$

(The case $0 = \theta_2(\xi_2,t) = \theta_{G2}(\xi_2,t) + \theta_{H1}(\xi_2,t)$ is considered in a similar way.) We recall that the unit eigenvector corresponding to $\theta_2(\xi_2,t)$ was denoted by $\mathbf{z}_2(\xi_2,t)$ in §6.3.3. By Theorem 6.3.1 it satisfies

$$\mathbf{z}_2(\xi_2,t) = \mathbf{z}_{G1}(\xi_2,t) \otimes \mathbf{z}_{H2}(\xi_2,t).$$

Therefore, the following theorem holds.

Theorem 6.4.1. *Let $t \in [0,1]$ be fixed. Further, let $\mu_2(t)$ be second smallest eigenvalue and $\mathbf{x}_2(t)$ be its corresponding eigenvector in the generalised eigenvalue problem*

$$\big[L_G \otimes I_n + I_m \otimes L_H\big]\mathbf{x}(t) = \mu(t)\big[D_G(t) \otimes I_n + I_m \otimes D_H(t)\big]\mathbf{x}(t),$$

where L_G and L_H are the Laplacian matrices of the weighted Path graphs G and H, respectively, $D_G(t) = (1-t)I_m + tD_G$ and $D_H(t) = (1-t)I_n + tD_H$. Finally, let

$$\theta_{G1}(t) < \theta_{G2}(t) < \cdots < \theta_{Gm}(t) \qquad \textit{and} \qquad \theta_{H1}(t) < \theta_{H2}(t) < \cdots < \theta_{Hn}(t)$$

be the eigenvalues of $L_G - \mu_2(t)D_G(t)$ and $L_H - \mu_2(t)D_H(t)$, respectively, and $\mathbf{z}_{Gi}(t)$ and $\mathbf{z}_{Hj}(t)$ be their corresponding unit eigenvectors, where $1 \leq i \leq m$ and $1 \leq j \leq n$.

Then, either

$$\theta_{G1}(t) + \theta_{H2}(t) = 0 \qquad \textit{and} \qquad \mathbf{x}_2(t) = \mathbf{z}_{G1}(t) \otimes \mathbf{z}_{H2}(t),$$

or

$$\theta_{G2}(t) + \theta_{H1}(t) = 0 \qquad \textit{and} \qquad \mathbf{x}_2(t) = \mathbf{z}_{G2}(t) \otimes \mathbf{z}_{H1}(t).$$

The following remark is the conclusion of this chapter.

Remark 6.4.1. *In fact, we know that* $\mathbf{x}_2(0)$ *is the Fiedler vector of the Laplacian matrix,*

$$L_{G\times H} = L_G \otimes I_n + I_m \otimes L_H,$$

and $\mathbf{x}_2(1)$ *is the normalised Fiedler vector of the normalised Laplacian matrix,*

$$\hat{L}_{G\times H} = D_{G\times H}^{-\frac{1}{2}} L_{G\times H} D_{G\times H}^{-\frac{1}{2}}.$$

Further, from Lemma 6.3.3 we have that in the Kronecker products $\mathbf{z}_{G1}(t) \otimes \mathbf{z}_{H2}(t)$ *and* $\mathbf{z}_{G2}(t) \otimes \mathbf{z}_{H1}(t)$ *the entries of the vectors* $\mathbf{z}_{G1}(t)$ *and* $\mathbf{z}_{H1}(t)$ *don't change sign and the entries of the vectors* $\mathbf{z}_{G2}(t)$ *and* $\mathbf{z}_{H2}(t)$ *change sign exactly once. Hence, if* $\mathbf{x}_2(t) = \mathbf{z}_{G1}(t) \otimes \mathbf{z}_{H2}(t)$ *for* $t = 0$ *or* $t = 1$*, and the change of sign in* $\mathbf{z}_{H2}(t)$ *occurs at the* $j_H(t)$*-th entry, then we will be splitting the graph* $G \times H$ *along the edges* $(u_i, v_{j_H(t)})$*,* $i = 1, 2, \ldots, m$*. Thus, we split the graph* $G \times H$ *"horizontally". In a similar way, if* $\mathbf{x}_2(t) = \mathbf{z}_{G2}(t) \otimes \mathbf{z}_{H1}(t)$ *(for* $t = 0$ *or* $t = 0$*) and* $\mathbf{z}_{G2}(t)$ *changes sign at the* $i_G(t)$*-th entry, then we cluster* $G \times H$ *by cutting it along the edges* $(u_{i_G(t)}, v_j)$*,* $j = 1, 2, \ldots, n$*. Or we say that we split the graph* $G \times H$ *"vertically".*

In the case when G *and* H *are unweighted Path graphs, we know from Corollary 6.3.2 that the* $i_G(t)$*-th and the* $j_H(t)$*-th entries in* $\mathbf{z}_{G2}(t)$ *and* $\mathbf{z}_{H2}(t)$ *respectively, are their "middle" entries. Therefore, if* G *and* H *are unweighted Path graphs, we split their product with respect to either of the Laplacian matrices,* $L_{G\times H}$ *or* $\hat{L}_{G\times H}$*, by cutting the graph "horizontally", or "vertically" through the middle.*

Chapter 7. Extensions and future work.

There are many natural extensions of the work in this thesis. These extensions cover a range of both theoretical questions and the application of the techniques described here to applications. We now list several areas for future work.

The work in §2 can be extended to an analytical way of computing the distribution of the largest eigenvalue of a Laplacian matrix. With the help of Conjecture 2.4.2, this may give a corresponding formula for the distribution of the 2-norm. Analysis of the spectrum of random Laplacian matrices, similar to that analysis of the spectrum of GOE matrices, will be useful in investigating the sensitivity of spectral clustering of networks. Also, the numerical tests of the three conjectures stated in §2.3 and §2.4 showed good agreement with the theory. Therefore, a proof of either of them will be helpful in understanding the similarities and differences between random symmetric matrices whose entries have a distribution symmetric with respect to zero. This will also broaden the class of matrices, by which we could model the errors in the data of networks, which could lead to models of the noise in the data which is more flexible than the present.

At the end of §3.7 we made a first step at discussing an application of the results of that section to the sensitivity to perturbation of the entries of eigenvectors. The analysis presented there needs to be refined and tested on real networks. In §3.8 we indicated briefly the extension of perturbation of eigenvalues and eigenvectors to perturbation of singular values and singular vectors. A natural next step is to complete the theory and then apply it to the clustering algorithm by (Higham et al., 2005). Further direction of the results presented in §3.7 and §3.8 could be their extension to measuring the sensitivity of subspaces to perturbation. Such an analysis could be useful when the spectral clustering is done by considering more than one eigenvector (or singular vector) of the matrix associated with the network.

As mentioned at the end of §4.3, the theory presented there does not have the limitations of the linearisation approach presented in §4.2, since the former is valid for any size of perturbation. However, it needs to be refined, to be less pessimistic, and then applied to micro-array data, for which the magnitude of uncertainty in the

data is known. When also extended to the sensitivity to perturbations of the entries of eigenvectors, or subspaces, the theory can be applied to measure the reliability of spectral clustering of noisy data, when the nature or magnitude of that noise is known.

The algorithm for discovering "master-slave" structures in networks of oscillators given in §5.4 needs to be extended and made rigorous for general networks. This can help by significantly reducing the computations for large networks of oscillators, in which only a few of the oscillators determine the dynamics of the rest.

The theory in §6 is a result of looking for any subtle agreements, or disagreements, between the spectral clustering produced by the Laplacian and normalised Laplacian matrices, associated with a given network. After failing to come to any conclusion in the general case of a weighted network (graph), we made a first step by the results given in that chapter. An extension to these results could lead to detecting (other) classes of graphs, for which the Laplacian and the normalised Laplacian matrices do, or alternatively, do not provide significant differences in the spectral clustering.

Chapter A. Appendix

A.1. Background Probability Theory

In this section we collect some standard results in Probability Theory, which we use throughout the thesis.

A.1.1. General Probability Theory

Definition A.1.1 (**σ- field**, c.f. (Grimmet and Stirzaker, 2001), p. 3)**.** Let Ω be a set. A collection $\mathcal{F}$ of subsets of Ω is called a σ-field if it satisfies the following conditions:

(a) $\emptyset \in \mathcal{F}$;

(b) if $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$;

(c) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

The elements of the σ-field $\mathcal{F}$ will be called **events**.

A set Ω with a σ-field $\mathcal{F}$ defined on it will be denoted as an ordered pair $(\Omega, \mathcal{F})$.

Definition A.1.2 (**Probability measure**, c.f. (Grimmet and Stirzaker, 2001), p. 5)**.** A *Probability measure* $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

(a) $\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\Omega] = 1$;

(b) if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, so that $A_i \cap A_j = \emptyset$ for all pairs i, j, satisfying $i \neq j$, then

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} A_i\right] = \sum_{n=1}^{\infty} \mathbb{P}[A_i].$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$, comprising a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

Definition A.1.3 (**Independent events**, c.f. (Grimmet and Stirzaker, 2001), p. 13)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The events $A, B \in \mathcal{F}$ are called **independent** if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

More generally, let I be an index set. The family of events $\{A_i\}_{i \in I}$ is called **independent** if

$$\mathbb{P}\Big(\bigcap_{i \in J} A_i\Big) = \prod_{i \in J} \mathbb{P}[A_i]$$

for all finite subsets J of I.

Definition A.1.4 (**Random variable**, c.f. (Grimmet and Stirzaker, 2001), p. 26)**.** A *random variable* is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega \mid X(\omega) < x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

Definition A.1.5 (**Independent random variables**, c.f. (Grimmet and Stirzaker, 2001), p. 91)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y : \Omega \to \mathbb{R}$ be two random variables on it. We say that X and Y are **independent** if the events

$$\{X < x\} \qquad \text{and} \qquad \{Y < y\}$$

are independent for all $x, y \in \mathbb{R}$ (see Definition A.1.3).

Example A.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable defined on it. Then it is easy to show that any constant $c \in \mathbb{R}$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Further, X and c are independent random variables. The latter can be shown by noting that*

$$\mathbb{P}[\{X < x\} \cap \{c < y\}] =: \mathbb{P}[X < x, c < y] = \begin{cases} \mathbb{P}[\{X < x\} \cap \emptyset] & \text{if } y \leq c \\ \mathbb{P}[\{X < x\} \cap \Omega] & \text{if } y > c, \end{cases}$$

which implies

$$\mathbb{P}[X < x, c < y] = \begin{cases} 0 & \text{if } y \leq c \\ \mathbb{P}[X < x] & \text{if } y > c. \end{cases}$$

Hence $\mathbb{P}[X < x, c < y] = \mathbb{P}[X < x]\mathbb{P}[c < y]$ for all $x, y \in \mathbb{R}$.

Definition A.1.6 (**Distribution function**, c.f. (Grimmet and Stirzaker, 2001), p. 26)**.** The *distribution function* of a random variable X is the function $F : \mathbb{R} \to [0, 1]$

given by[1]

$$F(x) = \mathbb{P}[X < x].$$

Remark A.1.1. Distribution functions *(see Definition A.1.6) are also called* cumulative distribution functions*, which we shall mostly use throughout this thesis. Sometimes we shall abbreviate* cumulative distribution function *by* c.d.f.*.*

Definition A.1.7 (**Discrete and continuous random variables, probability density function**, c.f. (Grimmet and Stirzaker, 2001), p. 32)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be some random variable defined on it.

(a) If $X(\Omega)$, that is, the image of X, is a countable set, then X is called a **discrete random variable**;

(b) If the distribution function of X, $F(x)$, can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)du \qquad \text{for all } x \in \mathbb{R}, \tag{A.1}$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$, then X is called a **continuous random variable** and the function f is called a **probability density function** (p.d.f.) of X.

Remark A.1.2. *The distribution function of a* continuous *random variable is certainly continuous (actually it is 'absolutely continuous') (c.f. (Grimmet and Stirzaker, 2001), p. 32). This is why* discrete *random variables cannot be* continuous*, and vice versa. However, there are random variables which are neither* discrete*, nor* continuous *(c.f. (c.f. (Grimmet and Stirzaker, 2001), p. 32).*

Remark A.1.3. *If the c.d.f. of a random variable, $F(x)$, is a differentiable function, then it is easy to show from* (A.1) *that $F'(x) = f(x)$.*

Definition A.1.8 (**Expectation**)**.** Let X be a random variable with cumulative distribution function $F(x)$. Then the *expectation* of X, $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] := \int_{-\infty}^{+\infty} x dF(x),$$

if the integral exists.

[1] To be precise, (Grimmet and Stirzaker, 2001) define $F(x) = \mathbb{P}[X \leq x]$ instead of $F(x) = \mathbb{P}[X < x]$, but since we consider random variables with continuous distribution functions, both approaches lead to the same results.

Theorem A.1.1 (**Basic properties of the expectation**, c.f. (Grimmet and Stirzaker, 2001), p. 52)**.** *The expectation operator* $\mathbb{E}$ *has the following properties:*

(a) If $X \geq 0$ *then* $\mathbb{E}[X] \geq 0$

(b) If $a, b \in \mathbb{R}$ *then* $\mathbb{E}[aX + bY] = a\mathbb{E}[x] + b\mathbb{E}[Y]$

(c) The random variable 1*, taking the value* 1 *always, has expectation* $\mathbb{E}[1] = 1$*.*

Lemma A.1.1 (**Independence and expectation**, c.f. (Grimmet and Stirzaker, 2001), p. 53)**.** *Let* X *and* Y *be independent random variables. Then* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$*.*

Definition A.1.9 (**Characteristic function**, c.f. (Grimmet and Stirzaker, 2001), p. 163)**.** Let X be a random variable. The *characteristic function* of X is the function $\varphi : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi(t) := \mathbb{E}[\exp(itX)],$$

where $i = \sqrt{-1}$

Characteric functions are a very powerful analytic tool for proving results in Probability Theory. In particular, the following theorem gives a way of calculating the *moments* of random variables from their *characteristic functions.*

Theorem A.1.2 (**Relation between characteristic functions and moments**)**.** *Let* X *be a random vriable and* $\varphi(t)$ *be its* characteristic function*. Then if the* n*-th moment of* X*,* $\mathbb{E}[X^n]$*, exists,* $\varphi(t)$ *can be differentiated* n *times at zero and*

$$\mathbb{E}[X^n] = i^{-n} \left[\frac{d^n}{dt^n} \varphi(t) \right]_{t=0},$$

where $i = \sqrt{-1}$*.*

Example A.1.2 (**Normal distribution**)**.** *The characteristic function of the normal distribution* $\mathcal{N}(\mu, \sigma^2)$ *is given by*

$$\varphi(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2} \right).$$

Therefore, if $X \in \mathcal{N}(0, 1)$*, the fourth moment of* X*,* $\mathbb{E}[X^4]$*, exists and is given by*

$$\mathbb{E}[X^4] = i^{-4} \left[\frac{d^4}{dt^4} \exp\left(-\frac{t^2}{2} \right) \right]_{t=0} = \left[(t^4 - 6t^2 + 3) \exp\left(-\frac{t^2}{2} \right) \right]_{t=0} = 3.$$

Definition A.1.10 (**Variance**)**.** Let X be a random variable with expectation $\mu = \mathbb{E}[X]$. Then the *variance* of X, $\mathrm{Var}[X]$, is defined as

$$\mathrm{Var}[X] := \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 dF(x),$$

if the integral exists.

Theorem A.1.3 (**Basic properties of the variance**, c.f. (Grimmet and Stirzaker, 2001), p. 53)**.** *If X and Y are independent random variables then*[2]

(a) $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$

(b) If $a \in \mathbb{R}$ *then* $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$

(c) $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Definition A.1.11 (**Normal distribution**)**.** We say that the random variable X has **normal distribution** with mean μ and variance σ^2, written $X \in \mathcal{N}(\mu, \sigma^2)$, if the cumulative distribution function of X is given by

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

Remark A.1.4. *From Definition A.1.11 we can see that every constant c can be regarded as a normally distributed random variable with mean c and variance* 0*, that is, $c \in \mathcal{N}(c, 0)$.*

Theorem A.1.4. *Let $X \in \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \in \mathcal{N}(\mu_Y, \sigma_Y^2)$ be two* independent *random variables and $a, b \in \mathbb{R}$ be some constants. Then $U := aX + bY$ is a normally distributed random variable with mean $\mu_U = a\mu_X + b\mu_Y$ and variance $\sigma_U^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$, that is, $U \in \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b\sigma_Y^2)$.*

Definition A.1.12 (*m*-**variate normal distribution**, c.f. (Muirhead, 1982), p. 5)**.** We say that the random vector $\mathbf{x} \in \mathbb{R}^{m\times 1}$ has an *m*-**variate normal distribution** if, for every deterministic vector $\boldsymbol{\alpha} \in \mathbb{R}^{m\times 1}$, the random variable $\boldsymbol{\alpha}^T\mathbf{x}$ has normal distribution.

And in particular:

Definition A.1.13 (**Joint normal distribution**)**.** Two random variables, x_1 and x_2, are said to have *joint normal distribution*, if the vector $\mathbf{x} = (x_1, x_2)^T$ has a bivariate normal distribution.

[2](Grimmet and Stirzaker, 2001) only state properties (b) and (c). We have included property (a) here for convenience.

Definition A.1.14 (**Bernoulli distribution**)**.** In general, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is said to have a **Bernoulli distribution** if its image, $X(\Omega)$, consists of only two values, say x_1 and x_2, and $\mathbb{P}[X = x_1] = p$ and $\mathbb{P}[X = x_2] = 1 - p$ for some $0 < p < 1$.

Definition A.1.15 (**Uniform distribution**)**.** A random variable X is said to have a **Uniform distribution** on the interval (a, b) for some $a < b$, if its distribution function, $F(x)$, is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b. \end{cases}$$

Remark A.1.5. *It is easy to show that the* probability density function *of X is then given by*

$$f(x) = \begin{cases} 0 & \textit{if } x \leq a \textit{ or } x \geq b \\ \dfrac{1}{b-a} & \textit{if } a < x < b. \end{cases}$$

A.1.2. Convergence of random variables

Definition A.1.16 (**Modes of convergence**, c.f. (Grimmet and Stirzaker, 2001), p. 274)**.** [3] Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that

(a) $X_n \to X$ *in r-th mean*, written $r \geq 1$, written $X_n \xrightarrow{r} X$, if $\mathbb{E}[|X_n^r|] < \infty$ for all $n \in \mathbb{N}$ and

$$\mathbb{E}[|X_n - X|^r] \to 0 \quad \text{as } n \to \infty;$$

(b) $X_n \to X$ *in probability*, written $X_n \xrightarrow{\mathbb{P}} X$, if

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| \geq \varepsilon] = 0 \qquad \text{for all } \varepsilon > 0;$$

(c) $X_n \to X$ *in distribution*, written $X_n \xrightarrow{\mathcal{D}} X$, if

$$\mathbb{P}[X_n < x] \to \mathbb{P}[X < x] \quad \text{as } n \to \infty$$

for all points x at which $F_X(x) = \mathbb{P}[X < x]$ is continuous.

[3] We have slightly modified the actual definition from (Grimmet and Stirzaker, 2001) by changing events of the form $\{X \leq x\}$ to $\{X < x\}$, but this is merely a matter of convention and both definitions are equivalent.

Theorem A.1.5 (**Relations between types of convergence**, c.f. (Grimmet and Stirzaker, 2001), p. 276)**.** *The following implications hold:*

$$(X_n \xrightarrow{r} X) \Rightarrow (X_n \xrightarrow{\mathbb{P}} X) \Rightarrow (X_n \xrightarrow{\mathcal{D}} X)$$

for any $r \geq 1$.

Theorem A.1.6 ((c.f. (Grimmet and Stirzaker, 2001), p. 277))**.** *If* $X_n \xrightarrow{\mathcal{D}} c$*, where* c *is constant, then* $X_n \xrightarrow{\mathbb{P}} c$.

Theorem A.1.7 (**Slutsky's theorem**, c.f. (Grimmet and Stirzaker, 2001), p. 285)**.** *Suppose that* $\{X_n\}_{n\in\mathbb{N}}$ *and* $\{Y_n\}_{n\in\mathbb{N}}$ *are two sequences of random variables satisfying* $X_n \xrightarrow{\mathcal{D}}$ *and* $Y_n \xrightarrow{\mathbb{P}} c$*, where* c *is a constant. Then* $X_nY_n \xrightarrow{\mathcal{D}} cX$ *and* $\frac{X_n}{Y_n} \xrightarrow{\mathcal{D}} \frac{X}{c}$ *if* $c \neq 0$*, and also* $X_n + Y_n \xrightarrow{\mathcal{D}} X + c$.[4]

Theorem A.1.8 (**Continuity theorem**, c.f. (Grimmet and Stirzaker, 2001), p. 172)**.** *Suppose that* $\{X_n\}_{n\in\mathbb{N}}$ *is a sequence of random variables with characteristic functions* $\varphi_1, \varphi_2, \ldots$.

(a) If $X_n \xrightarrow{\mathcal{D}} X$*, where* X *is a random variable with characteristic function* φ*, then* $\varphi_n(t) \to \varphi(t)$ *for all* t*;*

(b) Conversely, if $\varphi(t) = \lim_{n\to\infty} \varphi_n(t)$ *exists and is continuous at* $t = 0$*, then* φ *is a characteristic function of some random variable* X*, and* $X_n \xrightarrow{\mathcal{D}} X$.

Theorem A.1.9 (**Law of large numbers**)**.** *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of independent and identically distributed (i.i.d.) random variables with mean* $\mathbb{E}[X_i] = \mu < \infty$ *and variance* $\mathrm{Var}[X_i] = \sigma^2 < \infty$*. Then the sequence of random variables*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu, \tag{A.2}$$

as $n \to \infty$.

Remark A.1.6. *From Theorem A.1.5 it follows that Theorem A.1.9 is true if* $\xrightarrow{\mathbb{P}}$ *is replaced by* $\xrightarrow{\mathcal{D}}$ *in* (A.2).

Remark A.1.7. *Theorem A.1.9 is also known as* The Weak Law of Large Numbers*, because of the type of convergence in* (A.2)*. In comparison, in* The Strong Law of Large Numbers *the* convergence in probability *in* (A.2) *is replaced by* convergence almost surely (a.s.).

[4]The last part of this theorem can be found in (Billingsley, 1968).

Remark A.1.8. *In fact the requirement* $\mathrm{Var}[X_i] < \infty$ *in Theorem A.1.9 is* <u>not</u> *necessary. The theorem holds even if the variances of* X_i *are infinite, although the cnvergence is slower.*

A.2. MATLAB program which tests Conjecture 2.3.1

Program A.2.1 (MATLAB program which tests $G_n(t) \stackrel{n}{=} F_n(t)^2$ (see Conjecture 2.3.1 and (2.20))).

```
n = 500; % The dimension of the matrix B we simulate.
nrep = 1e4; % Number of simulations.

% Initialisation of MaxEig and NormB for better performance.
MinEig = zeros(nrep, 1);
MaxEig = zeros(nrep, 1);
NormB = zeros(nrep, 1);

% The simulations.
matlabpool open 2
parfor ii = 1:nrep
    B = randn(n);
    B = (B + B')/2; % B is from GOE.
    spect = eig(B);
    MinEig(ii) = min(spect);
    MaxEig(ii) = max(spect);
    NormB(ii) = max(abs(MinEig(ii)), abs(MaxEig(ii)));
end
matlabpool close

% The cumulative distribution functions of MaxEig and
% NormB, respectively, based on the simulations.
% Here f = F_n(t) and g = G_n(t).
[f, x] = ecdf(MaxEig);
[g, s] = ecdf(NormB);

% Comparison between the plots of f^2 and g, that is,
% between F_n(t)^2 and G_n(t).
```

```
figure
stairs(x, f.^2, '--k')
hold on
stairs(s, g, ':k')
title(['n = ', num2str(n)])
xlabel('t')
ylabel('G_n^{(S)}(t)', 'rotation', 0)
h = legend('F_n^{(S)}(t)^2', 'G_n^{(S)}(t)');
set(h, 'Location', 'Best')
hold off
```

Bibliography

Z. D. Bai and Y. Q. Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix. *Ann. Prob.* **16**, 1988.

A-L. Barabasi. *Linked: How everything is connected to everything else and what it means for business and everyday life.* Plume Books, 2003.

P. Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, 1968.

B. Bollobas. *Random Graphs.* Academic Press, New York, 1995.

M. J. Bowick and E. Brezin. Universal scaling of the tail of the density of eigenvalues in random matrix models. *Phys. Letts.* **B268**, pages 21–28, 1991.

S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a web in your pocket? *Data Engineering Bulletin*, 21:37–47, 1998.

F. Chung. *Spectral Graph Theory.* American Mathematical Society, 1997.

C.H.Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE International Conference on Data Mining, 2001. ICDM 2001*, pages 107–114, 2001.

W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

A. Edelman and P-O. Persson. Numerical methods for eigenvalue distributions of random matrices, 2005. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:math-ph/0501068`.

M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory. *Czech. Math. Journal*, 25:619–633, 1975.

M. Fiedler. Algebraic connectivity of graphs. *Czech. Math. Journal*, 23:298–305, 1973.

A. S. Fokas, A. R. Its, and A. V. Kitaev. The isomonodromy approach to matric models in 2d quantum gravity. *Communications in Mathematical Physics*, 147(2):395–430, July 1992.

P. J. Forrester. The spectrum of the edge of random matrix ensembles. *Nucl. Phys.* **B402**[FS], pages 702–728, 1993.

F. R. Gantmacher. *Matrix Theory.* Chelsea Publishing Company, New York, 1959.

G. Grimmet and D. Stirzaker. *Probability and Random Processes.* Oxford, 3rd ed. edition, 2001.

P. Grindrod. Range dependent random graphs and their application to modelling small world proteomic data sets. *Phys. Rev. E*, 66, 2002.

S. Guattery and G. L. Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19:701–719, 1998.

D. J. Higham, G. Kalna, and J. K. Vass. Analysis of the singular value decomposition as a tool for processing microarray expression data. *Algoritmy*, pages 250–259, 2005.

D. J. Higham, G. Kalna, and M. Kibble. Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics*, 204(1), 2007.

R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, 1990.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

Y. Kuramoto. Self-entrainment of a population of coupled nonlinear oscillators. In H. Araki, editor, *International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Physics*, volume 39, pages 420–422. Springer, New York, 1975.

Y. Kuramoto. *Chemical Oscillators, Waves, and Turbulence.* Springer Verlag, Heidelberg, 1984.

C. McDuffee. *Theory of matrices.* 1946.

M. L. Mehta. *Random Matrices.* Academic, San Diego, 1991.

R. J. Muirhead. *Aspects of multivariate statistical theory.* John Wiley & Sons, 1982.

S. Nagaraj, S. Bates, and C. Schlegel. Application of eigenspace analysis techniques to ad-hoc networks. In *Lecture Notes in Computer Science*, volume 3158, pages 300–305. Springer-Verlag, 2004.

M. E. J. Newman. Modularity and community structure in networks. *PROC.NATL.ACAD.SCI.USA*, 103:8577, 2006. URL `doi:10.1073/pnas.0601602103`.

M. E. J. Newman, B. Karrer, and E. Levina. Robustness of community structure in networks. *Physical Review E*, 77:046119, 2008. URL `doi:10.1103/PhysRevE.77.046119`.

B. N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, 1998.

A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with eigenvectors of graph. *SIAM Journal of Matrix Anal. Appl.*, 11:430–452, 1990.

W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850, Dec. 1971.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 2000.

D. Spielman. Graph theory and its applications. (online lecture notes), 2004.

G. W. Stewart. Stochastic perturbation theory. *SIAM Review*, 32(4):579–610, 1990.

C. A. Tracy and H. Widom. The distribution of the largest eigenvalue of the gaussian ensembles: $\beta = 1, 2, 4$. *in Calogero-Moser-Sutherland Models, eds. J.F. van Diejen and L. Vinet, CRM Series in Mathematical Physics 4, Springer-Verlag, New York*, 2000a.

C. A. Tracy and H. Widom. Airy kernel and Painlevé II. 2000b.

C. A. Tracy and H. Widom. Level spacing distributions and the airy kernel. *Commun. Math. Phys.* **159**, 1994a.

C. A. Tracy and H. Widom. Fredholm determinants, differential equations and matrix models. *Commun. Math. Phys.* **163**, 1994b.

C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.*, 177:727–754, 1996.

J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, 1965.

W. Yueh. Eigenvalues of several tridiagonal matrices. *Applied Mathematics E-Notes*, 5:66–74, 2005.